TEXT MINING THE BIOMEDICAL LITERATURE

By

Dr. Ronald N. Kostoff
Office of Naval Research
875 N. Randolph St.
Arlington, VA 22217
Phone: 703-696-4198
Internet: kostofr@onr.navy.mil

(The views in this report are solely those of the author, and do not represent the views of the Department of the Navy or any of its components)

KEYWORDS

Text Mining; Information Retrieval; Metrics; Bibliometrics; Computational Linguistics; Biomedical; Information Technology; Asymmetry Detection; Citation Analysis; Literature-Based Discovery; Literature-Assisted Discovery.

ABSTRACT

Text mining of the biomedical literature provides patterns of relationships among concepts, people, and institutions, offering enhanced medical/ technical intelligence unobtainable by other means. This report describes myriad text mining capabilities. It starts with a description of the larger context of knowledge management, then, addresses components of text mining in particular.

Section 1 covers biomedical knowledge management, the role of text mining in knowledge management, and describes the cultural changes and global agreements required to allow the full power and capabilities of text mining to be utilized. The next two sections address information retrieval issues. Section 2 describes the extraction of useful information from the published biomedical literature. Section 3 describes the information content in different record fields, in a major medical database.

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
05 NOV 2007		**00-00-2007 to 00-00-2007**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Text Mining the Biomedical Literature	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Office of Naval Research,Dr. Ronald N. Kostoff,875 N. Randolph St.,Arlington,VA,22217**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Text mining of the biomedical literature provides patterns of relationships among concepts, people, and institutions, offering enhanced medical/ technical intelligence unobtainable by other means. This report describes myriad text mining capabilities. It starts with a description of the larger context of knowledge management, then, addresses components of text mining in particular. Section 1 covers biomedical knowledge management, the role of text mining in knowledge management, and describes the cultural changes and global agreements required to allow the full power and capabilities of text mining to be utilized. The next two sections address information retrieval issues. Section 2 describes the extraction of useful information from the published biomedical literature. Section 3 describes the information content in different record fields, in a major medical database.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**	**Same as Report (SAR)**	**380**	

The next four sections address computational linguistics issues, especially related to identifying patterns and relationships in text. Section 4 outlines a family of methods for generating radical biomedical discovery from the literature. An ongoing application of literature-based discovery to treatments for Raynaud's Disease shows that one to two orders of magnitude more discovery than competing literature-based techniques are possible. Section 5 shows how increasing specialization within the biomedical community creates roadblocks for the acceleration of radical discovery described in Section 4, and recommends ways to eliminate these roadblocks. Section 6 describes the detection of unexpected asymmetries from the biomedical literature, with a specific example on bilateral organ cancer incidence asymmetry detection. Section 7 describes a unique approach for removing words/ phrases of low technical content, and improving the quality of the resulting technical taxonomies.

The final six sections address the use of citation analysis in biomedical text mining. Section 8 describes the use and misuse of citation analysis in biomedical text mining. Section 9 describes Citation Mining, a technique that allows the myriad impacts of research to be identified and assessed. Section 10 describes the use of citation analysis to evaluate the quality of research performers. Section 11 shows a systematic approach for defining the seminal literature of any biomedical topic, critical for literature reviews or background sections of papers. Sections 12 and 13 describe the differences between highly and poorly cited biomedical articles, with specific case studies from leading medical journals. Finally, a comprehensive bibliography on text mining, along with associated infrastructure bibliometrics, is included.

Section 1. Biomedical Knowledge Management

(based on Kostoff, R. N. "Science and Technology Knowledge Management". in New Frontiers of Knowledge Management. (Ed.) Kevin DeSouza. Palgrave Macmillan, United Kingdom. 11-35. 2005.)

OVERVIEW

This section focuses on knowledge management in the biomedical science and technology (S&T) development cycle. It has three major components: textual data documentation and dissemination; the conversion of textual data into knowledge; and the impact of literature-based knowledge on the biomedical S&T development cycle. For each component, the author's version of ideal operation is outlined, a contrast is drawn to present-day operation, and recommendations for closing the gaps are presented. The operational details are based on the author's experiences with text processing and mining. The constraints imposed by core textual data deficiencies on the processing results and value are emphasized, and methods for reducing data deficiencies are presented.

INTRODUCTION

Modern economies and militaries depend on continual advances in science and technology (S&T) for competitive advantage. Optimal S&T progress depends on adequate resources, strategic planning, an environment conducive to discovery and innovation, a technically literate workforce, and an awareness of the total range of S&T being conducted globally. This awareness accelerates S&T progress when the knowledge available from global S&T research is integrated into the strategic and tactical decision-making involved in initiating, managing, and applying S&T. Any reluctance to share information by an individual, organization, or country not only impacts the global S&T community, but impacts the reluctant unit negatively as well. Performance and credibility are required to get a 'seat at the S&T table', which translates to people must 'share information to receive an information share'. This need for information sharing must be balanced with information security requirements. At the level of the firm, competitive advantage maintenance dictates security considerations. At the national level, national security must be maintained.

At the core of this knowledge to accelerate research is the assemblage of documents known collectively as the technical literature. This literature is

the foundation of S&T; it is the 'glue' that binds S&T vertically and horizontally, across disciplines, cultures, and time. The building blocks of this literature, the technical documents themselves, constitute a strategic resource of the Information Age. The manipulations of these blocks, especially their analysis and synthesis, become a critical activity of the Knowledge Age. The short-term products of these manipulations, the immediate decisions and actions resulting from the analysis and synthesis, become tactical outputs of the Knowledge Age, while the longer-term impacts of the decisions and actions become strategic outcomes from the Knowledge Age.

The impact of knowledge on decision-making for biomedical S&T development will be no better than the adequacy and quality of the foundational technical documents, the conversion to knowledge of the textual data embodied in these documents, and the integration of this knowledge into the management of the biomedical S&T development cycle. This section will address these three critical components: textual data documentation and dissemination; conversion of textual data to knowledge; integration of technical knowledge into the biomedical S&T development cycle. For each component, the section will: 1) present a vision of ideal operating conditions; 2) contrast this with the operating conditions present today; and 3) provide recommendations for closing the gap between the actual and the ideal. First, some definitions are required.

DEFINITIONS

Data are the transduced outputs of sensors. For published technical text, data are the text representations of inputs to the human mind from experiments, theoretical studies, computer studies, and internally-generated ideas. The human mind becomes the transducer, deciding what data to represent as text, and how the data will be represented. *Information* is the fusion of data (the creation of the network that 'connects the dots' that data represent), and incorporates both the data and the relationships among data. For textual data, information incorporates the patterns and quantitative relationships among words, phrases, and grammatical structures. *Knowledge* is the placement of information in its larger context. It is the necessary condition for understanding; without context, there is little understanding. For technical text, knowledge allows the text patterns and quantitative relationships to be interpreted within the context of the over-arching technical issues.

Data mining is the extraction of useful information and knowledge from raw data. Unlike physical mining, data mining does not deplete the ore, but rather enriches it with new derived data to mine. Presently, the information extraction is performed by a combination of machines and humans. Text *mining* is the text analog of data mining, the extraction of useful information and knowledge from the literature. Many documents are typically used in text mining, and are subjected to rigorous analytical processes. Therefore, the results are typically more objective than the limited experiences and biases of a few people analyzing a few documents, characteristic of pre/non-text mining approaches. *S&T text mining* is the extraction of useful technical information and knowledge from the technical literature, and is the key component of the conversion of technical data to technical knowledge. *S&T knowledge management* is the use of global S&T data to accelerate progress in S&T development and enhance decision-making relative to S&T.

As an example of these concepts, consider the technical discipline of nanotechnology, the study and exploitation of phenomena at the nanometer scale (atomic level). Suppose a researcher wants to understand the structure of a material surface at the atomic scale level. The researcher might use an Atomic Force Microscope to measure surface topography. The signals from the photo-diode detector represent raw data related to atomic surface topography, and when quantified and recorded, represent raw numerical data. When fused with calibration data and analyzed, information on surface topography is generated. The information results from observing patterns in, and relationships among, the separate data elements. When limited to a small region of the surface, this information is rudimentary. As the region of measurement increases, the information becomes more comprehensive. When this comprehensive information is analyzed in the context of theories and experiments using other techniques and other materials, knowledge is generated.

When this research is documented, S&T text mining can be applied. The individual words and phrases are the raw textual data, the relationships among these textual data elements represent information, and the placing of this information in context of other technical studies becomes knowledge. While text mining can be performed on single documents, especially large documents, in the modern day context it is usually performed on aggregates of documents. If individual nanotechnology documents are assembled into a

large database, then text mining of the total nanotechnology database operates on a much larger amount of textual data. It can produce substantially more information because of the larger number, and greater diversity, of textual data points and the resultant larger number of patterns generated. Greater knowledge can then be produced because of the wider scope of data analyzed, greater complexity of relationships, more data points supporting statistical analyses, and increased predictive validity due to improved statistics.

CRITICAL COMPONENTS OF BIOMEDICAL S&T KNOWLEDGE MANAGEMENT

The biomedical S&T knowledge management cycle can be viewed as consisting of three major steps, and sub-steps within some of the steps:

1. **Technical Textual Data Documentation and Dissemination**
    a. Textual Data Documentation
    b. Textual Data Publication
    c. Textual Data Dissemination
2. **Conversion of Technical Textual Data to Technical Knowledge**
    a. Data Retrieval
    b. Data Processing-Information Generation
    c. Information Integration-Knowledge Generation
3. **Integration of Technical Knowledge into the S&T Development Cycle**

For each sub-step, the format used will be similar. Initially, the author's version of the ideal sub-step process is presented (e.g., the first sub-step would start with ideal textual data documentation), followed by the actual sub-step process (e.g., actual textual data documentation, for the first sub-step), and ending with recommendations for closing the gap between the ideal and the actual.

Technical Textual Data Documentation and Dissemination

This step spans the gamut from documenting the initial text describing the S&T actually performed to distributing this textual data to processors and analysts. It represents the critical path to knowledge generation. The most sophisticated text analysis algorithms cannot

compensate for raw data deficiencies, whether due to non-existence of this data, or non-availability.

Textual Data Documentation

The foundational step of the technical literature is the creation of its component documents. Ideally, all S&T performed globally would be documented, including detailed descriptions of technical successes, failures, and promising research terminated prematurely. Having all these data available would allow strategic and tactical S&T development decisions to be made with knowledge of all past and present S&T (and future plans as well), thereby profiting from other researchers' successes, as well as from their failures.

While a reasonable fraction of S&T performed is probably documented in some form (notes, memos, records), only a small fraction is documented in a form targeted for wide distribution (Kostoff, 2003a). There are many disincentives to publish in widely accessible literatures (classification, proprietary, time tradeoff between publishing and doing actual research, etc), and few incentives to publish. The majority of S&T researchers who publish, especially in basic science, tend to represent segments of academia that reward publications. Thus, much of the S&T actually performed is unavailable to the majority of global researchers, providing an intrinsic constraint on the efficiency and quality of S&T development.

A fundamental assumption of the corrective recommendations to follow is that those who pay the bills for S&T have the responsibility to insure its quality. The first step to insuring S&T quality is becoming aware of the S&T produced with the funds provided. In part, S&T awareness requires documentation of S&T by its producers, and gaining access to this documented S&T by its sponsors.

There are two major sponsors of S&T globally: government, and industry. Government is almost the exclusive sponsor of higher-risk science today, especially basic science, while industry plays a large role in sponsoring lower-risk technology development. While there are relatively few incentives and motivations available to increase open literature publication from industry (other than patents), there are many available to government.

The first, and main, recommendation of this chapter is that the sovereign governments of the world should take steps to come to agreement that publication of all publicly-sponsored research (with exceptions for classified and highly sensitive research) in a form conducive for wide-scale dissemination has value for all participants. Then, through a series of mandates and incentives, the governments should take steps to insure that documentation of publicly-funded research occurs in readily accessible media. The challenge here is to have governments accept the concept that the documentation and wide-scale dissemination of research have comparable importance to the research itself for the advancement and utilization of S&T, and are in fact integral components of the conduct of research. As a starting point, consensus needs to be gained among the agencies of the U. S. Federal government that 'publication of all publicly-sponsored research (with exceptions for classified and highly sensitive research) in a form conducive for wide-scale dissemination has value for all participants'.

Textual Data Publication

In an ideal publication environment, the initial documentation of research in all forms (notes, memos, records, e-mail, etc) should be required to enter the more formal technical literature. This is the starting point for making the documentation accessible to a wider audience. Because of the peer review process, this step maintains a threshold of quality for the research documentation. Second, in this ideal publication environment, all journals would have the same formatting requirements. Time would not have to be expended re-writing text and references to match unique journal requirements. Third, the review process and period would be uniform across journals.

Fourth, reviewers would be compensated by the journals. Funds to underwrite the review process would come from the Federal agencies, with some fraction going to the journal, and the remainder going to the reviewers. A necessary condition for journals to receive these Federal funds is that they adhere to the uniform formatting and review process standards. Because of compensation, the reviewers would be expected to devote more time to the review, thereby generating a more extensive and credible review. In turn, a ceiling would be placed on the review period, and the reviewers would get paid proportionally less as the time ceiling is exceeded.

Fifth, there would be uniform review criteria for all articles of a similar type, and authors would understand what issues to address before manuscript submission. Sixth, all components of the technical paper would be structured, in the sense that canonical criteria for each technical paper component would have to be addressed. Structured Abstracts have been used by many medical journals for over a decade (Haynes, 1990; Kostoff and Hartley, 2002a), and a body of knowledge exists to provide guidance for future structuring. Similar structuring should be required for Titles, Keywords (Hartley and Kostoff, 2003), and References (Kostoff and Hartley, 2003c).

Finally, for researchers who submit a manuscript and are employed by an organization, the results of the manuscript review would be sent to both the employing organization and the research sponsor. This latter scenario is a double-edged sword. From the positive perspective, it would serve as a high quality project review, as well as to minimize authors' shopping around for journals in which to publish mediocre papers. From the negative perspective, it could increase reluctance of researchers to publish, if they know their management will see the reviewers' comments.

Would any of these ideal publication requirements, especially those relating to uniformity of formatting, structure, and evaluation criteria, stifle innovation and creativity in the manuscript? Uniformity in formatting would certainly not, and would in fact leave more time for creativity by requiring less time to be spent on re-formatting documents, especially references. Uniform structuring would be a double-edged sword. For some authors, more time would be required, to incorporate canonical issues that would not have been addressed by the author without the requirement. On the other hand, having to address these additional structural criteria would (at a minimum) make a paper more complete, and could stimulate additional creativity by having the author address difficult issues that would otherwise have been finessed and avoided. In summary, having uniformity of formatting and canonical submission/ review requirements should have little, if any, impact on creativity, but much impact on availability of substantially more technical information to the larger technical community.

Contrast this ideal publication process with technical publication today. First, there are no requirements to publish research results in the more formal technical literature, and few incentives and motivations to do so for much of the S&T community. Second, most journals have different

formatting requirements, and substantial time is expended to re-write text, especially the references. Third, the review process and review period can be vastly different across journals, and the review period can vary widely within the same journal, depending on the review managing editor and the reviewers selected. The present author has had papers accepted after reviews ranging from a week after manuscript submission to over eighteen months beyond manuscript submission. There is no excuse for a lengthy drawn-out review process, and at the pace of change of some of today's research, the results could be obsolete by the time the research is published.

Fourth, reviewers serve gratis today, as part of professional responsibility. This usually translates to a lower priority for the review, relative to more lucrative tasks. As a result, many reviewers devote small amounts of time to manuscripts that embody hundreds or thousands of hours of work. The lowered priority limits the accuracy, timeliness, and benefits of the review. Fifth, most journals don't supply detailed review criteria to the authors, and the author has a relatively vague understanding of what is expected before manuscript submission. Additionally, the criteria supplied to the reviewers are different for most journals, sometimes vastly different.

Sixth, most journals don't require any structuring of the research paper, other than section formatting. As a result, the levels of information provided in a research paper, and especially in the most widely-read component, the Abstract, are very different. Many medical journals now require Structured Abstracts (i.e., authors should address background, objectives, approach, results, conclusions, etc), but not structuring of any other fields. For electronic retrieval, different search engines access different fields, and it is now important that different fields contain similar information at different levels of resolution (Kostoff and Hartley, 2003c).

Finally, manuscript reviews today are sent to the author(s) only. This has two negative consequences. Authors can shop around for journals, go through the submission and review process until they find a favorable journal, and the failed reviews that highlight a paper's deficiencies will remain unknown. Also, a major opportunity for program review is missed. Many research programs do not get anywhere near the level of detailed review that exists in the journal manuscript review process, if in fact these programs get reviewed at all. Having the journal manuscript reviewers serve as the proxy program reviewers would insert a much higher level of quality control (at least for the purely technical component of the program) than

exists in many programs today. Obtaining this higher level of program quality control, as well as making the research documentation available to a much wider technical community, are the two main reasons for having Federal funding of the research reviews.

Improvement in publication practices to approach the ideal vision above requires the agreement of individual journal editors and their publishers, agreement of the professional societies to which many of the journals belong, and especially agreement among the sovereign nations of the world to contribute funding for support of the journal manuscript review process. These changes could be implemented through top-down dictums from the sovereign governments. However, a more realistic first step is to gain consensus from the technical journal editors that the ideal publication vision presented above will benefit S&T, the nations, and especially the journals themselves. How can this consensus be obtained?

An important goal of most journals is to make money, in addition to providing a forum for research progress in particular disciplines. A key feature of the ideal publication scenario above is government subsidies of the journals through compensating the journals and manuscript reviewers for each manuscript submitted. To prevent 'flooding' of the journals with frivolous submissions, some modest contribution from the authors toward the review process may be required as well.

Under this scenario, two ways for a journal to increase revenues are increasing the volume of submissions, and increasing the throughput of reviews by more efficient processes. If the journal editors could be convinced that standardizing and simplifying many aspects of the manuscript submission and review process would encourage more researchers to submit their manuscripts for publication, and would increase the throughput of manuscripts, they would welcome these changes as sources of increased revenue. Once a critical mass of journals, especially the influential journals (those with high Impact Factors, high circulation, and high absolute number of citations), is convinced these changes are beneficial, it would only be a matter of time before the remainder of journals agrees as well.

Textual Data Dissemination

Ideal dissemination is a function of technical literature type. In its broader aspects, the technical literature encompasses at least the following types of documentation: peer-reviewed journal articles; magazine articles; conference proceedings; organizational technical reports; books; patents; research proposals; program/ project narrative descriptions. With the exception of books, most of the other types of technical literature are becoming available electronically. Since the electronic forms are most amenable to large-scale text mining, their dissemination electronically is the focal point of this section.

First, the existence of these large electronic databases would be advertised widely in the technical community, to make potential users aware of their contents, uses, and benefits. Second, the scope and content of these databases would be heavily influenced by two major communities: the performer-user community (the researchers who use these databases to improve the quality and conduct of their research), and the sponsor-user community (those sponsors and evaluators who use these databases to 1) track the productivity and impact characteristics of the sponsored research to insure the research funds are being spent effectively and efficiently, and 2) input the research results into the planning and selecting of future research). For journal article and conference proceedings databases especially, these user communities would determine the journals to be covered and the record fields to be included.

Third, the interfaces with these databases would be standardized, so the users would not be required to learn a new query language whenever they used a different database. Fourth, the capabilities of these database search engines would incorporate the latest technology available, and would be standardized across databases. These multi-database search capabilities are starting to appear in very limited cases (for example, the NSF, NIH, and DOE awards databases accessible through one search engine at www.osti.gov/fedrnd). Fifth, many of the costs for accessing these databases would be subsidized by the Federal government S&T sponsors, to encourage more widespread use.

Sixth, those agency-specific technical databases, such as research proposals or research program/ project narrative descriptions, would be shared with other government agencies, at a minimum. For those agency databases that do not contain agency-sensitive information, the contents would be shared with the public. Finally, strict quality control would exist. This includes

completeness (records would not be accepted by the large databases unless they contained all fields); currency (database contents updated periodically, obsolete records eliminated); and accountability (contents of each record approved by responsible individual).

There are serious deficiencies in the dissemination of all of the above types of technical literature, and they severely restrict the information available to the decision-makers on S&T development. First, there is very little awareness of the existence of these large electronic databases within the technical community. Relatively few technical people outside the text mining community use these databases on a regular basis. Second, the scope and content of the major electronic databases, especially the commercial databases, are determined by the database developers, not the users. The organizations that mainly sponsor the S&T (upon which these databases draw) have little impact on their contents.

Third, the search engine interfaces are different among the databases, and users effectively have to learn a new search engine query language when they encounter a new database. This is particularly difficult for people who use the databases sporadically, since they may not be willing to invest the time to learn these new languages for every database use. The breadth of interfaces is unnecessary, and serves as a barrier to wider-scale use.

Fourth, the technology capabilities vary strongly across databases. For example, the Medline search engine PubMed allows 10,000 records to be downloaded at once from the Web version, whereas the Science Citation Index search engine allows only 500 records total to be downloaded from any Web-based search. As another example, some databases contain sponsor information, others do not. Fifth, the costs for accessing many of these databases are prohibitive. High costs effectively filter out competent analysts who do not have large grants that include database access costs, or do not work for organizations that have contractual access to these databases. Since many of these databases are one-of-a-kind, the developers effectively control monopolies, and impose monopoly pricing. Paradoxically, data, a strategic resource of the Information Age, has become monopolized!

At the same time, use of these large technical databases is becoming more critical in the technological component of the war against terrorism. These databases provide linkages among people, institutions, and technologies.

Disappearance of names or technologies from the technical literature over time might have significance for evolution of technologies into negative applications. Integrating individual narrowly-focused S&T thrusts over an institution or country could provide insights into total systems capabilities being developed, not evident at the single project level. Tracking the evolution of co-authors, or secondary or tertiary links to co-authors, and developing a network of associations, could provide a larger picture of unhealthy relationships not evident at the individual scientist level. For these, and many other reasons, allowing the technical database strategic resource to become monopolized is unacceptable.

Sixth, there is a real problem in the sharing of agency technical databases, such as research proposals or program/ project descriptive narratives. Lack of detailed awareness of other agency programs (or proposals) translates directly into lack of coordination and effective use of resources. This could result in 1) multiple agencies pursuing similar work, or 2) agencies not being aware of prior research activities, and repeating past mistakes, or 3) agencies not participating in joint efforts that would fully exploit each agency's strengths.

Finally, quality control of many of these databases leaves much to be desired. Not all records in the journal article/ conference proceedings databases, for example, contain Abstracts. Not all records contain Keywords. Some database providers function as warehouses. In those cases, whatever the database distributor receives from the source is entered into the database.

The main dissemination problems identified above stem from the multiplicity of database owners/ developers, and a lack of a unified set of standards the databases must meet. It is difficult to envision how a uniform set of standards could be implemented and enforced across the myriad technical databases that exist today, with the present form of database ownership. The National Institutes of Health (NIH)/ National Library of Medicine (NLM) sponsorship of PubMed/ Medline should be the model for database ownership, although certainly some of the deficiencies mentioned above apply in part to PubMed/ Medline as well.

Due to the strategic resource nature of the technical databases, especially in their potential relationship to the war on terrorism in particular and national defense in general, they should be transferred to specific Federal agencies

for development and maintenance. Funds would be added to these agencies' budgets for this purpose. For example, NIH/ NLM manage PubMed/ Medline, which is appropriate considering their mission. The National Science Foundation (NSF) could manage the basic/ early applied research database, and National Institute of Standards and Technology (NIST) could manage the late applied/ technology development/ engineering database. The United States Patent and Trademark Office (USPTO) could manage the Patent database, as is presently the case, but it would have to be upgraded substantially to support large volume patent analyses.

While these databases are presently separated, and proposed as separated above, there are many cases where linkages among databases would be valuable. If all the databases were to contain references, this would be one source of valuable linkages, especially for studying the problem of conversion of science to technology. Although separate, these databases do have some overlap. For example, the two databases SCI and Medline share common medical journals, and SCI and EC share common engineering science journals. When quantitative studies of technical categories are being performed for specific literatures, duplication of journals and associated double-counting can lead to significant errors. Perhaps search engines that link to multiple databases, and have the capability to eliminate duplicate records, would be desirable.

To eliminate the agency 'stove-piping' problem (where database sharing is less than ideal), a centralized capability that stores, or at least accesses, these multiple agency databases would be required. As an example of one of the benefits of this arrangement, all S&T proposals would be accessible by at least Federal agency personnel through this centralized process. There would have to be oversight for quality control, perhaps by Office of Science and Technology Policy (OSTP), independent of whether the databases are supplied to some central facility or whether they are retained by the sponsoring agency, with the capability for remote access. Many implementation issues would need to be addressed, to insure protection of privacy, proprietary information, classified information, currency of material, and cost-sharing.

Conversion of Technical Textual Data to Technical Knowledge

This step constitutes what is usually perceived as text mining. It ranges from data retrieval to the generating of information by processing, and finally to knowledge generation by the integration of information. While there is wide-scale belief that this step is almost completely automated due to the prevalence of sophisticated text analysis software, in reality high quality knowledge generation requires high quality human analysts working with high quality analytic processes.

Data Retrieval

The objectives of textual data retrieval, usually known by the misnomer 'information retrieval', are twofold: retrieve all documents of interest in the technical literature examined (known as 'recall'), and retrieve as few non-relevant documents as possible (precision). The common wisdom is that a tradeoff exists between precision and recall. However, use of relevance feedback approaches, where linguistic patterns are identified in relevant and non-relevant documents, and used in an iterative query development process, can provide both high precision and high recall. Ideal textual data retrieval consists of high recall with high precision, with minimal human reading of documents to judge relevance, minimal human intervention to identify patterns, relatively few iterations, and relatively short retrieval times.

The need for Internet search engines that are simple, efficient, and precise has spurred much research in textual data retrieval. As a result, fairly sophisticated algorithms exist to provide high quality textual data retrieval. Many of these methods have been reported in the proceedings of the annual TREC conference (e.g., TREC, 2003), hosted by NIST, and in the major journals Information Processing and Management, Journal of the American Society for Information Science and Technology, and Journal of Information Science.

Unfortunately, few people do serious textual data retrieval at all. Additionally, the number of people in the technical community who do retrieval and use these advanced algorithms, especially within the context of advanced retrieval processes, is miniscule. The author did a review of biomedical literature search techniques (actually used and reported by researchers in the literature) a few years ago, as part of a large study on

biomedical information retrieval (Kostoff, 2001b). Most of the research articles examined, even those in which state-of-the-art literature reviews were conducted, reported very simplistic queries, consisting of very few terms, and not imbedded within any type of sophisticated retrieval process. There is a major gap between what retrieval technology has to offer, and what is used by the technical community in practice.

Two major steps are needed to improve retrieval: encourage researchers and other S&T users to do electronic textual data retrieval more often, and train them in the latest software and retrieval processes. Some government agencies mandate the former as a required step before initiation of research, but these mandates are rarely enforced. The mandates need to be extended across all agencies, and seriously enforced. The latter retrieval training should be provided to all S&T personnel, both performers and sponsors, with training updated and refreshed periodically.

Data Processing – Information Generation

The core of textual data processing consists of bibliometrics and computational linguistics. Bibliometrics is applicable to databases with large numbers of records, each of which has attributes such as author, address, source, references, etc. Evaluative bibliometrics (Narin, 1976; Garfield, 1985; Schubert et al, 1987) uses counts of patents, publications, citations, and other potentially informative items to develop science and technology performance indicators. Its validity is based on the following premises: counts of patents and papers are a valid indicator of R&D activity in the subject area of those patents or papers; the number of times those patents or papers are cited in subsequent patents or papers is a valid indicator of the importance or impact of the cited patent or paper; and, the citations from papers to papers, from patents to patents, from papers to patents, and from patents to papers are an indicator of the intellectual linkages between the organizations that are producing the patents and papers, and knowledge linkages between the subject areas (Narin et al, 1994).

Ideally, all databases containing these types of records would include populated fields for all these attributes. The attribute entries would be standardized, such that attribute values that are identical in reality would be represented as identical in the database. In particular, there would be a unique representation for each unique author, even multiple authors with the same name. Different references for the same journal would be converted to

one standardized representation. Different representations of the same organizational address would be converted to one standardized representation. For databases of science and technology articles, references would be represented with the same level of detail as the articles themselves.

Citation-Assisted Background (CAB) is a bibliometrics-based systematic approach to identifying the seminal papers in a technical discipline (Kostoff and Shlesinger, 2004c). After the modern literature in a technical discipline has been retrieved, all the references are extracted, and the most highly-cited references in different time periods are identified. The thrusts of, and links among, these references are described, to generate either the Background section of a research paper or a literature review/ survey. The identification of seminal papers reflects the collective viewpoint of the larger technical community, not the experiences or biases of a few individuals.

Ideally, CAB would be sufficient to identify all the seminal documents in a discipline. To insure adequate representation from the major sub-disciplines, the retrieved literature would be clustered, and CAB would be applied to each cluster as well as to the total literature. Additionally, the highly-cited documents would themselves be clustered, to help structure/ categorize the Background section/ literature review document.

S&T computational linguistics (Mitkov, 2002) is a statistical process for quantifying text patterns that underlies the extraction of useful information from large volumes of technical text. It identifies pervasive technical themes in large databases from technical phrases that occur frequently. It also identifies relationships among these themes by grouping (clustering) these phrases (or their parent documents) on the basis of similarity. Further, it uncovers theme relationships by correlating phrase occurrences and subsequently identifying pervasive factors in the database through factor analysis. Ideally, all non-technical clutter would be removed from the extracted textual data before the pattern analysis is initiated. All the technical phrases would be selected automatically.

Clustering (Rasmussen, 1992) is a sub-set of computational linguistics, and groups similar objects. It has two major types: concept and document. Concept clustering (Kostoff et al, 2002b) groups words/ phrases to form thematic categories and taxonomies. Document clustering (Steinbach et al, 2000; Willet, 1988) groups the documents that contain these words/ phrases to form categories and taxonomies as well. The main advantage of

document clustering is that the number of documents in each category is identified, providing some indication of the levels of effort/ emphasis in each technical category. If the analyst is interested in assessing adequacy or deficiency of effort in each category, then the allocation of documents across categories (from which other attributes, such as funding, can be derived) provides an excellent starting point for such computations.

Ideal clustering removes trivial words and phrases automatically, such that grouping can occur based on technical text similarities, rather than on non-technical concepts. It assigns single items to multiple clusters where appropriate, such that a multi-theme text is represented in all relevant technical clusters. Ideal clustering does not require exact word matching for grouping, but allows similar concepts to be combined. It provides hierarchical taxonomies as well as flat taxonomies, since homogeneous databases (e.g., single technologies with complementary sub-technologies) are amenable to hierarchical taxonomies whereas heterogeneous databases (e.g., multiple disparate technologies) are more amenable to flat taxonomies. Finally, ideal clustering provides category attributes (e.g., funding, organizations, etc) at all hierarchical levels.

Factor analysis (Jackson, 1991) aims to reduce the number of variables in a system, and to detect structure in the relationships among variables. Inter-variable correlations are computed, and highly correlated groups (factors) are identified. The relationships of these variables to the resultant factors are displayed clearly in the factor matrix, whose rows are variables and columns are factors.

Ideally, the number of factors would be selected automatically. The phrases input to the factor analysis would be conflated automatically, and different phrases reflecting the same concept would be conflated automatically as well. An algorithm would generate a hierarchical tree reflecting a hierarchical taxonomy whose categories are the aggregated factors.

Citation Mining (Kostoff et al, 2001a) and Reference Mining are two specific examples of the more generic trans-citation mining, where analysis done on one side of the citation boundary can yield information about the other side. In Citation Mining, it is desired to identify characteristics about the users of research, or more specifically, the sub-set of users who cite the research of interest. The documents output by a research unit (e.g., a person, research group, organization, nation, discipline) are examined, and the

documents that cite them are retrieved. These retrieved citing documents are text mined, and the characteristics of the citing documents are related to those of the cited documents. For Reference Mining, substitute 'reference' for 'citation'.

Since Citation Mining is text mining applied to citations, the ideal conditions for Citation Mining correlate with those for text mining listed in the previous sections. In addition, multiple generations of citations (i.e., Citation Mining of citations) would be desired for research impact studies, and these multiple generations would be easily retrievable in the ideal situation.

Literature-Based Discovery (LBD) is a method of generating new knowledge and discovery based on literature analysis alone (Swanson, 1986; Hearst, 1999; Kostoff, 2003b). The end result is hypotheses, to be tested by further experiment. From one perspective, LBD is the fourth branch of research, following theory, experiment, and computer data modeling.

The leading demonstrated LBD approach (Swanson and Smalheiser, 1997) is based on the premise that specialization of today's researchers prevents the interdisciplinary knowledge that many problems require for solution. This LBD approach links these disjoint literatures (through intermediate literatures directly linked to each) to identify concepts in one discipline that could be of value to solving problems in other disciplines. As an example of how this specific LBD approach works, consider its first published example, the treatment of Raynaud's Disease (Swanson, 1986). Papers addressing Raynaud's Disease, otherwise called the Raynaud's Disease (RD) literature, were retrieved from a medical database. Major themes from the RD literature, such as blood viscosity (BV), were extracted by text mining, and examined further. The BV literature was then retrieved, and its major themes were examined. Those themes in the BV literature present in the RD literature were eliminated, and the remaining themes were considered as candidates for discovery. Fish Oil and Eicosapentanoic Acid were viewed as the most promising, and published in the literature. Subsequent lab experiments and clinical tests verified the hypothesis.

Ideally, the major LBD themes would be extracted automatically. An algorithm would be available to identify the most promising of these themes. In the subsequent retrieved literatures, the important concepts would be identified, and those concepts (not only words or phrases) that are disjoint from the originating literature, and are thus candidates for discovery, would

be identified. An algorithm would be available for prioritizing the potential candidates in terms of importance.

Asymmetry detection identifies asymmetries that are unexpected, by extracting patterns from text, and evaluating them for symmetry (Goldman et al, 1999). For example, a recent paper detected bilateral asymmetries of cancer incidence in human organs from text mining statistics of large numbers of published narratives in the Medline database (Kostoff, 2003e). These asymmetries would not ordinarily be expected. In the above paper, asymmetries were detected in such terms as left lung carcinoma vs. right lung carcinoma, or left kidney carcinoma vs. right kidney carcinoma, and were validated against actual patient incidence data.

Ideally, an algorithm would be available to identify such asymmetries. Additionally, asymmetries could be detected for different levels of detail (e.g., left vs. right lung cancer, left lower lobe vs. right lower lobe lung cancer, etc).

In actual practice, many of the problems that impact bibliometrics stem from the existence of multiple databases developed by multiple organizations with no standardizing requirements. In any database, only selected attributes are represented, and the attributes differ for different databases. For any attribute, not all records will be populated, and the attribute formats will differ across databases.

A serious problem is the lack of ability of any databases to distinguish among different authors with the same name. For bibliometrics studies of large databases, this can lead to substantial occurrence frequency inaccuracies with respect to common names. Even in the same database, the same organization and organizational unit are referenced differently by different authors, reducing the actual occurrence frequencies of the organization.

In actual computational linguistics, perhaps the two major problems are the large amount of clutter (context-dependent trivial phrases, and the difficulty in removing them without removing important technical phrases) and the different textual representations of the same concept. Clustering in particular is seriously impacted by large clutter, and clusters can be driven by the clutter rather than the high technical content phrases.

In most clustering algorithms, one record or one phrase is assigned to one cluster. For multi-theme records, the other themes are then effectively disregarded. Finally, the different textual representations of the same concept (mentioned above) lead to fragmented clusters, except in some isolated cases where a thesaurus can be used to standardize representations.

In actual factor analysis, selection of the number of factors to analyze is not straight-forward. Multiple approaches exist (e.g., Scree Plots (Cattell, 1966), eigenvalue thresholds (Kaiser, 1960)), and each can give radically different values for numbers of clusters. Phrase conflation (combining words/ phrases with same stems), a context-dependent phenomenon (Kostoff, 2003d), is usually performed with context-independent stemming algorithms (Porter, 1980), and usually requires manual supplementation for conflating acronyms and synonyms. Finally, mapping from factors to a hierarchical taxonomy is usually a manual process (Kostoff, 2002b).

Citation-Assisted Background is a recently-developed process (Kostoff and Shlesinger, 2004c), and so far has identified seminal papers based on citations from the total retrieved database as a unit. In theory, it is possible that some sub-themes could be under-represented in seminal documents unless citations from each cluster were assessed. Actual Citation Mining has the same limitations as those listed for text mining above. Additionally, selecting generations of citations (e.g., citations of cited papers, citations of citations of cited papers, etc) with the available databases and search engines is a manually intensive process, and can very rapidly lead to selection of overwhelming and infeasible numbers of documents.

The main limitations with Literature-Based Discovery are the large numbers of potentially-related themes, and the much larger number of candidate discovery concepts possible from each theme. For all practical purposes, these candidate discovery concepts must be sifted and prioritized manually. Finally, while Asymmetry Detection may have reasonable accuracy at high-level relatively standard representations (e.g., left vs. right, high vs. low), accuracy drops sharply at more detailed levels of representation because of the different ways in which concepts are represented in text (e.g., left lower lung, lower left lung, left lower lobe, etc).

The first and foremost recommendation for improving information generation is that the data deficiencies identified in the previous sections must be addressed. Second, improved techniques for removing clutter in a

context-dependent manner must be developed. Third, clustering algorithms need to be improved to allow objects to be placed in multiple clusters, and to allow different textual representations of the same concept to be clustered together.

Information-Integration-Knowledge Generation

Evaluative bibliometrics can be used: (i) to identify the infrastructure (authors, journals, institutions) of a technical domain (Kostoff et al, 2000a, 2004b); (ii) to identify experts for innovation-enhancing technical workshops and review panels; (iii) to develop site visitation strategies for assessment of prolific organizations globally; (iv) to identify impacts (literature citations) of individuals, research units, organizations, and countries (Kostoff et al, 2000b, 2004a), and; (v) to identify the seminal documents underlying a technical discipline (Kostoff and Shlesinger, 2004c), based on the larger community's perception of the importance of these documents.

Computational linguistics (including its sub-categories described above) can be used for: (i) enhancing information retrieval and increasing awareness of the global technical literature (Kostoff et al, 1997; Greengrass, 1997; TREC, 2003); (ii) discovery and innovation based on merging common linkages among very disparate literatures (Swanson, 1986; Swanson, 1997; Kostoff, 2003b; Gordon, 1998); (iii) uncovering unexpected asymmetries in the technical literature (Goldman et al, 1999; Kostoff, 2003e); (iv) estimating global levels of effort in S&T sub-disciplines (Kostoff et al, 2000b; 2004b; Viator, 2001); (v) helping authors to increase their citation statistics by improving access to their published papers, which may help journals increase their Impact Factors (Kostoff et al, 2004a, 2004b); and (vi) tracking research impacts across time and applications areas (Davidse, 1997; Kostoff et al, 2001a).

Applications of evaluative bibliometrics are limited by the data quality deficiencies described previously, especially the absence of attribute quality standards and the absence of citation data in most large databases. To expand the bibliometrics applications, the underlying data deficiencies must be eliminated.

Applications of the computational linguistics results (including those of all the computational linguistics sub-categories described above) are limited by

data deficiencies described previously, and the lack of processes for interpreting and exploiting clustering results. Most of the research effort in this area has been on algorithm development, as opposed to process development. As a result, excellent clustering tools have been developed, without the parallel understanding of how they should be integrated into the larger evaluation context.

The most challenging of the computational linguistics applications, as well as the application with the highest potential payoff, is Literature-Based Discovery. Unfortunately, it has received almost negligible support, and has effectively fallen between the cracks. It is extremely difficult to do, since systematic processes have not yet been validated. It is one of the few truly inter-disciplinary research areas, and suffers from the dis-incentives common to inter-disciplinary research (difficult to obtain funding, difficult to publish results, difficult to gain external awards). Its need for complex process development runs counter to the present day mainline information technology community trend of developing algorithms for licensing and sale.

Three major steps are required to improve integration. First, address the data deficiencies, as described previously. In parallel, expand the access to databases, especially other government databases, to expand the breadth of applications. Second, expend more resources to support clustering analytical process development, to bring tool development and process development into balance. Third, initiate a serious national program on Literature-Based Discovery.

Integration of Technical Knowledge into the S&T Development Cycle

S&T development can be divided into a number of phases, including S&T planning, identification and selection of new S&T, execution and review of S&T, and technology transition/ transfer. Each of these phases requires knowledge of global S&T activity to maximally accelerate the development of S&T. Some uses of this knowledge in each of the development phases are discussed here.

Planning is the cornerstone of S&T development, and is perhaps the phase that makes/ could make optimal use of global S&T knowledge. Many organizations include visitation of S&T institutions globally (global site visitations) as part of their technology assessment function. Bibliometrics is useful for planning a global site visitation strategy, in order to insure that the

critical producers and institutions are examined. Bibliometrics identifies the prolific producers and centers of excellence, and many other infrastructure attributes of interest. Bibliometrics is also helpful for identifying expert participants for advisory planning panels, and experts for planning workshops as well.

Clustering and taxonomy generation are useful for categorizing outputs of a technical discipline globally, and then assessing the adequacy and deficiencies of S&T investments in each category (Kostoff et al, 2000b). Clustering and taxonomy generation have also been used by the author to categorize a technical discipline's literature into thematic groups, and provide the categorized literature as background material for panels and workshops.

Citation-Assisted Background (CAB) is ideal for structuring a comprehensive literature review before initiation of a project, in the planning stage. Because it shows the many roots that feed a discipline, it has the potential for providing insights to new research directions.

Additionally, CAB has the potential for Literature-Based Discovery. The leading LBD approach today (Swanson and Smalheiser, 1997) uses literatures that are indirectly related to a primary literature through intermediate directly-related literatures to search for discovery. In the previous Raynaud's Disease example shown for LBD, the Blood Viscosity literature was directly related to both the Raynaud's Disease literature and the Fish Oil literature, and was used to link the indirectly related Fish Oil and Raynaud's Disease literatures. These can be viewed as spatially related concepts, since the linkage is through co-occurrence of text elements in the retrieved literatures.

If the CAB process were applied to the clusters of the retrieved literature, then the cited literatures would become analogous to the indirectly related literatures above. Thus, if one of the clusters of the Raynaud's Disease literature focused on Blood Viscosity, its citations would constitute a literature (or literatures) indirectly related to the Raynaud's Disease literature. If Fish Oil were a sub-set of the cited literature, then the discovery would be through the citation route rather than the co-occurrence route. The CAB LBD approach is based on temporally-related concepts (citations reflect previous time), as opposed to the present mainline approach's being based on spatially-related concepts. Eventual integration

of the two approaches could search for discovery across space and time, and expand the discovery space substantially.

Literature-Based Discovery can identify the multiple disciplines required to solve complex problems, as well as identifying the experts in those disciplines. It has the potential for identifying new research directions and promising opportunities (Swanson, 1986; Swanson and Smalheiser, 1997), as well as for structuring the agenda for innovation workshops and identifying workshop participants (Kostoff, 2003b).

Identifying and selecting the S&T to be funded, especially for basic research, is a most critical activity. Most agencies use independent panels or mail reviewers, and bibliometrics is extremely useful for identifying experts to fill these roles. Literature-Based Discovery can be used, in some cases, to solve specific problems, and to propose hypotheses that specific research programs will address.

In the execution, and particularly the review, phases (See Appendix for an illustration of text mining support of S&T review processes), placing periodic progress review results in the context of global S&T development is a key to insuring cutting-edge research. Bibliometrics aids in selecting the review experts, and in providing metrics to gauge the productivity and impact of the research. Information retrieval and clustering provide the global literature context and levels of effort.

Transition/transfer is perhaps the most difficult phase of the process, since it involves bridging the gap between two communities, research and advanced development. There are two major components of the technology transition/ transfer process. The first is the potential user becoming aware of the S&T, and the second is the potential user taking specific steps to develop further/ implement the S&T. While knowledge management is limited in its contribution to the second component, it has the potential for major contributions to the first component. Ideally, transition planning is initiated at the start of the research process, potential users/ customers are identified, and these users/ customers are made active participants of advisory panels for the research project. Information retrieval, and especially Citation Mining, plays important roles in identifying potential customers for the research.

REFERENCES FOR SECTION 1

Cattell, R.B. (1966). The Scree Test for the number of factors. Multivariate Behavioral Research. 1. 245 –276.

Davidse RJ, Van Raan AFJ. (1997). Out of particles: impact of CERN, DESY, and SLAC research to fields other than physics. Scientometrics. 40:2 . 171-193.

Garfield E. (1985). History of citation indexes for chemistry - a brief review. JCICS. 25(3). 170-174.

Goldman JA, Chu, WW, Parker, DS, Goldman, RM. (1999). Term domain distribution analysis: a data mining tool for text databases. Methods of Information in Medicine. 38. 96-101.

Gordon MD, Dumais S. (1998). Using latent semantic indexing for literature based discovery. Journal of the American Society for Information Science. 49 (8): 674-685.

Greengrass E. (1997). Information retrieval: An overview. National Security Agency. TR-R52-02-96.

Hartley, J. and Kostoff, R. N. (2003). How useful are 'key words' in scientific journals?" Journal of Information Science. 29:5. 433-438. October.

Haynes, R.B., Mulrow, C. D., Huth, E. J., et al, (1990). More informative abstracts revisited. Ann. Intern. Med. 113: 69-76

Hearst MA. (1999). Untangling text data mining. Proceedings of ACL 99, the 37th Annual Meeting of the Association for Computational Linguistics, University of Maryland, June 20-26.

Jackson, J. E. (1991). A users guide to principal components. Wiley, New York, p. 569.

Kaiser, H.F. (1960). The application of electronic computers to factor analysis. Educational and psychological measurement. 20: 141-151.

Kostoff RN, Eberhart HJ, Toothman DR. (1997). Database Tomography for information retrieval. Journal of Information Science. 23:4.

Kostoff RN, Braun T, Schubert A, Toothman DR, and Humenik JA. (2000a). Fullerene roadmaps using bibliometrics and Database Tomography. Journal of Chemical Information and Computer Science. 40(1): 19-39.

Kostoff, R. N., Green, K. A., Toothman, D. R., and Humenik, J. (2000b). Database tomography applied to an aircraft science and technology investment strategy". Journal of Aircraft, 37:4. 727-730. July-August.

Kostoff RN, Del Rio JA, García EO, Ramírez AM, Humenik JA. (2001a). Citation mining: integrating text mining and bibliometrics for research

user profiling. Journal of the American Society for Information Science and Technology. 52:13. 1148-1156.
Kostoff, R. N. (2001b). The extraction of useful information from the biomedical literature". Academic Medicine. 76:12. December.
Kostoff, R. N., and Schaller, R. R. (2001c). Science and technology roadmaps. IEEE Transactions on Engineering Management. 48:2. 132-143. May.
Kostoff, R. N., and Hartley J. (2002a). Structured abstracts for technical journals. Journal of Information Science. 28:3. 257-261.
Kostoff, R. N., Tshiteya, R., Pfeil, K. M., and Humenik, J. A. (2002b). Electrochemical power source roadmaps using bibliometrics and database tomography. Journal of Power Sources. 110:1. 163-176.
Kostoff, R. N. (2003a). Text mining for global technology watch. In Encyclopedia of Library and Information Science, Second Edition. Drake, M., Ed. Marcel Dekker, Inc. New York, NY. Vol. 4. 2789-2799.
Kostoff RN. (2003b). Stimulating innovation. International Handbook of Innovation. Larisa V. Shavinina (ed.). Elsevier Social and Behavioral Sciences, Oxford, UK.
Kostoff, R. N., and Hartley, J. (2003c). "Science and technology text mining: structured papers". DTIC Technical Report Number ADA417220.
Kostoff, R. N. (2003d). The practice and malpractice of stemming. JASIST. 54:10. 984-985. August.
Kostoff, RN. (2003e). Bilateral asymmetry prediction. Medical Hypotheses. 61:2. 265-266.
Kostoff RN, Shlesinger M, Malpohl G. (2004a). Fractals roadmaps using bibliometrics and database tomography. Fractals. 12:1. March.
Kostoff RN, Shlesinger M, Tshiteya R. (2004b). Nonlinear dynamics roadmaps using bibliometrics and Database Tomography. International Journal of Bifurcation and Chaos. 14:1. 61-92.
Kostoff, RN. and Shlesinger, MF(2004c). CAB – Citation-assisted background. Scientometrics. In Press.
Kostoff, RN., Karpouzian, G., and Malpohl, G. (2004d). Abrupt wing stall roadmaps using database tomography and bibliometrics". Journal of Aircraft. In Press.
Mitkov, R. (2002). The Oxford handbook of computational linguistics. Mitkov, R. (ed). Oxford University Press. UK.
Narin F, Olivastro D, Stevens KA. (1994). Bibliometrics theory, practice and problems. Evaluation Review. 18(1). 65-76.

Narin F. (1976). Evaluative bibliometrics: the use of publication and citation analysis in the evaluation of scientific activity (monograph). NSF C-637. National Science Foundation. Contract NSF C-627. NTIS Accession No. PB252339/AS.
Porter, M.F. (1980). An algorithm for suffix stripping. Program, 14(3). 130-137.
Rasmussen E. (1992). Clustering algorithms. In W. B. Frakes and R. Baeza-Yates (eds.). Information Retrieval Data Structures and Algorithms. Prentice Hall, N. J.
Schubert A, Glanzel W, Braun T. (1987). Subject field characteristic citation scores and scales for assessing research performance. Scientometrics. 12 (5-6): 267-291.
Steinbach M, Karypis G, Kumar V. (2000). A comparison of document clustering techniques. Technical Report #00--034. 2000. Department of Computer Science and Engineering. University of Minnesota.
Swanson DR, Smalheiser NR. (1997). An interactive system for finding complementary literatures: a stimulus to scientific discovery. Artificial Intelligence, 91:2. 183-203.
Swanson DR. (1986). Fish oil, Raynauds Syndrome, and undiscovered public knowledge. Perspectives in Biology and Medicine. .30: (1). 7-18.
TREC (2003). (Text Retrieval Conference), Home Page, http://trec.nist.gov/.
Viator JA, Pestorius FM. (2001). Investigating trends in acoustics research from 1970-1999. Journal of the Acoustical Society of America. 109 (5): 1779-1783 Part 1.
Willet P. (1988). Recent trends in hierarchical document clustering: A critical review. Information Processing and Management. 24:577-597.

APPENDIX: ILLUSTRATION OF TEXT MINING TO SUPPORT S&T PROGRAM REVIEW

In an S&T program review, three main questions are asked: Is the S&T program doing the right job? Is the S&T program doing the job right? Is the S&T program performing?

How can text mining help generate answers to each of these questions?

Is the S&T program doing the right job?

In order of precedence, this is the first issue to address. It focuses on the adequacy of the existing S&T investment strategy and associated roadmaps.

It starts with a vision, or description of the operational scenario. This is followed by an elucidation of the capabilities required in order for the vision to be implemented. The capabilities are quantified to provide the development targets. A roadmap of the S&T required to achieve the targets is generated, in parallel with the associated (ideal) investment strategy. This strategy consists of the investment allocation, and the rationale that supports the allocation.

The investment strategy can also be viewed as consisting of investment principles, investment allocations, and the investment rationale. Again, the actual can be compared against the ideal. What are some of these investment principles? Following are some of the investment principles used by the author in S&T investment strategy assessments:

- Is the balance among technical thrust areas appropriate?
- Is the balance among mission areas appropriate?
- Is the balance among funding categories (basic research, applied research, technology development) appropriate?
- Is the balance between discretionary and non-discretionary funding appropriate?
- Is the balance between 'technology push' and 'requirements pull' appropriate?
- Is the balance between revolutionary and evolutionary research appropriate?
- Is the balance between technology advancement and demonstration appropriate?
- Is the balance between high risk and low risk research appropriate?
- Is the balance among short term, intermediate term, and long term research appropriate?
- Is the balance between new projects and continuing projects appropriate?
- Is the balance among performers (university/ government/ industry) appropriate?
- Is the balance between individual research and joint projects (multi-department, multi-agency, multi-national, and government-industry) appropriate?
- Is the balance among single discipline, multiple discipline, and interdisciplinary research appropriate?
- Is the balance between large and small projects appropriate?

- Is the balance among research products (hardware, software, patents, presentations, reports, peer-reviewed journal papers) appropriate?

Obviously, additional investment principles are possible, depending on the specific review objectives, and review management interests. Once the desired S&T direction has been established, then the existing S&T program investment strategy is compared against the ideal investment strategy. Deviations of the existing from the ideal are noted, discussed, and corrective actions are taken, including personnel and budgetary.

Text mining could be used to support identification of the capabilities needed to implement the vision, development of the roadmap components, assessment of how well the investment principles are being followed, and evaluation of how the actual investment allocations compare with the desired investment allocations.

Support Identification of Capabilities

This would use the techniques employed, for example, in an Aircraft Investment Strategy study (Kostoff et al, 2000b). The evaluators would gather a number of different requirements documents, perform phrase frequency and proximity analyses, and identify technical capabilities to be pursued. The evaluators would then add planning documents, perform similar analyses, and identify enabling technologies for those capabilities.

Develop Roadmap Components

A roadmap is a network of technologies linked over space and time, aimed at achieving specific goals (Kostoff and Schaller, 2001c). A prospective roadmap is a network of science and technology areas to be developed in order to achieve the goals. Key issues in roadmap development center about whether all the blocks have been identified, all the linkages have been identified, and how accurate are the linkage strength quantifications.

Block identification comprehensiveness is a measure of how well the roadmap developers understand the mixture of technologies required to produce the higher level capabilities, and are aware of global S&T development. Advanced information retrieval, and associated clustering, can provide the mixture of technologies required to achieve the desired

capabilities. Advanced information retrieval can certainly identify relevant S&T being developed globally.

Linkage identification is a measure of how well the roadmap developers understand the relationships among the roadmap technologies. Proximity and co-occurrence analyses, performed on a database of technology narratives, should be able to provide the connections.

Linkage strength quantifications measure how well the roadmap developers understand the strength of the relationships. Again, phrase proximity analyses, which provide the co-occurrence frequencies of specific phrases (number of times phrase pairs co-occur in the same linguistic domain- e.g., paper Abstract), should be able to estimate these relationship strengths.

Assess Adherence to Investment Principles

The combination of clustering and bibliometrics would address the relationship between the actual investment allocations, and the ideal. Clustering groups documents (or words/ phrases) into categories, and if the core documents have associated attributes (funding, performers, institutions), then the weighted attributes in each category can be determined. Bibliometrics tend to count semi-structured data (authors, institutions, journals, countries) in each category.

Compare Actual Investment Allocations with Desired Investment Allocations

This is similar to what was done in the Aircraft investment strategy paper, although the far more powerful document clustering techniques developed recently (Kostoff et al, 2004d) could be used.

Is the S&T program doing the job right?

In order of precedence, this is the second issue to address. It focuses on the accuracy and efficiency of achieving the specified technical target. It evaluates the mechanics of the S&T development approach, and incorporates the cost, performance, schedule, and risk aspects of the mechanics. Most reviews concentrate on this component. Text mining could examine all the high frequency phrases, and all the cluster categories/ themes. Then, judgments would be made as to balance (e.g., too much theory relative to experiment, insufficient North American contributions,

etc). Examination of cluster themes and technical phrases has been done in almost every one of the author's technology text mining studies, has been validated with world-class experts in those disciplines, and has been shown to be a remarkably accurate indicator of deficiencies in specific technologies.

Is the S&T program performing?

There are three components of performance: productivity, impact, and progress. Here, text mining can be very helpful, depending on the metrics selected. Bibliometrics can provide information relative to publications, patents, and citations, where the publications and patents are productivity metrics, and the citations are impact metrics. Citation Mining, a combination of text mining and citation analysis, can provide impacts and audience accessed for a research unit. Progress, in the present context, addresses how well a program is meeting its technology readiness levels, or milestone targets. The relation of text mining to progress assessment is untested and not clear at this point.

Section 2. The Extraction of Useful Information from the BioMedical Literature.

(based on Kostoff, R. N. "The Extraction of Useful Information from the BioMedical Literature". Academic Medicine. 76:12. December 2001.)

OVERVIEW

Modern information technology provides the biomedical professional with powerful tools and processes for extracting useful information from large volumes of text. Presently, very little use is made of the full capabilities of these information technology tools to supplement research and teaching. The purpose of this section is to overview these tools and processes, and show the diversity of ways they can be applied to enhance the capabilities of biomedical professionals. The information technology terms are defined, requirements for high quality information extraction are presented, some available techniques for high quality information extraction are described/ referenced, and myriad information extraction applications are summarized. While substantial benefits from high quality information extraction techniques are possible, substantial time and effort, and technical expertise, are required to generate a credible high quality product.

BACKGROUND

Data mining, the extraction of useful information from large volumes of data, has become a useful approach for identifying important patterns buried in massive data sets. An important evolving sub-set of data mining is text mining, the extraction of useful information from large volumes of text. This section will focus on the use of text mining to extract relevant information from the biomedical and related literatures. After the components and importance of text mining are described, the section will present requirements for high quality information extraction from text, will overview/ reference some leading techniques used to extract this information, and will then conclude by summarizing diverse applications of text mining.

There are three major components of S&T text mining.

1) Information Retrieval
2) Information Processing
3) Information Integration

Information retrieval is the selection of documents or text segments from source text databases for further processing. Information processing is the application of bibliometric and computational linguistics and clustering techniques to the retrieved text to typically provide ordering, classification, and quantification to the formerly unstructured material. Information integration combines the computer output with the human cognitive processes to produce a greater understanding of the technical areas of interest.

High quality text mining applied to the global science and technology (S&T) literature can:

1) Enhance the retrieval of information from global S&T databases;
2) Identify the technology infrastructure (authors, journals, organizations) of a technical area;
3) Discover new technical concepts or new technical relationships from related or disparate technical literatures;
4) Identify and categorize the main technical themes and sub-themes in a large body of technical literature;
5) Identify the relationships between technical themes and infrastructure components;
6) Estimate global levels of emphasis in technical areas or sub-areas; use these results as the basis for S&T adequacy or deficiency judgements;
7) Provide roadmaps of myriad research impacts.

High quality S&T text mining can be a valuable adjunct for the biomedical researcher or clinician, in accessing and understanding increasingly larger amounts of available biomedical information and supporting S&T from other physical and engineering science branches. However, high quality text mining is not an automatic process. It requires time, effort, and above all, the combined use of technical experts in the focal area of interest, in related technical areas, and in information technology for maximum effectiveness.

WHAT IS THE PURPOSE OF THIS SECTION?

This section shows how the use of modern information technology, specifically text mining, can assist the biomedical researcher and clinician in accessing and understanding the relevant global S&T literature. Since present-day biomedical research and clinical practice have very high

technology components, the relevant global S&T literature encompasses not only biomedicine, but topics from the physical and engineering sciences as well.

For example, advances in biomedical instrumentation require underlying advances in materials, electronics, signal processing, mathematical analysis, physics, chemistry, energy conversion, radiation sciences, solid and fluid mechanics, robotics and micro-technology, and other technologies depending on specific applications. Maximum advances in non-invasive medical diagnostics require access to the latest science and engineering literature in remote sensing, non-destructive evaluation, signal and image processing, pattern recognition, multi-source data fusion, fluid dynamics, acoustics, robotics, materials, electronics, and many other disciplines.

This section is aimed primarily at the biomedical researcher and clinician. However, it should be of substantial value to anyone involved with biomedical S&T, including research managers, evaluators, administrators, and sponsors, as well as corporate and national security intelligence analysts and biomedical system investors.

WHY IS ACCESSING AND UNDERSTANDING THE GLOBAL RELEVANT S&T LITERATURE IMPORTANT?

Science and technology form the core of modern economies and militaries. Global S&T expenditures are in the neighborhood of $500-800B annually, depending on one's definition of S&T. No single organization, or even nation, can begin to cover the full spectrum of S&T development required for a modern competitive economy or military, due to resource limitations. Maximum efficiency in S&T resource expenditures begins with maximum exploitation of external S&T resources. This requires leveraging through cooperative S&T planning and development efforts, based on maximal awareness of external S&T efforts from overt and covert intelligence efforts.

For the biomedical researcher and clinician, maintaining awareness of the most recent research and technology development for practical purposes requires knowing what information is available and where it is stored, how to access it, and how to process and interpret it for maximal information content. Any lack of global S&T awareness can waste time, people, funding, and other resources due to:

1) Pursuing approaches demonstrated to be non-productive
2) Duplicating previously performed S&T
3) Not using the latest techniques and instrumentation
4) Pursuing research in the absence of guidance available from the literature
5) Making research decisions based on incomplete information

WHY ARE PRESENT INFORMATION RETRIEVAL AND ANALYSIS APPROACHES INSUFFICIENT?

An ongoing study illuminates the limitations of existing information retrieval practices. As part of a recent assessment of information retrieval techniques, I examined many biomedical studies that included literature searches. The Science Citation Index (SCI) Abstracts of these studies contained the queries used for the literature surveys. These queries had the following characteristics:

1) The source data came almost exclusively from Medline alone, except for those studies whose objective was to survey the Web resources available for the target medical issue;
2) Most of the studies focused on narrowly defined medical problems, with little indication offered that supporting or related medical/ technical areas were of any interest;
3) The reported queries contained 3-6 phrases on average;
4) The phrases were either searcher-generated, or were the indexed terms from the Medline Mesh taxonomy. No evidence was presented that an exhaustive search of author-generated terms was performed.

In my experience, queries with the above characteristics result in a deficient retrieved information base. These deficiencies translate into limitations on the credibility and quality of study results and subsequent research and development (R&D), for the following reasons.

1) Searches that do not access the myriad databases available, and queries that do not result in comprehensive retrievals of the information available in the databases actually searched, result in only a fraction of the existing knowledge being available for study and R&D exploitation.
2) Searches and queries not designed to a) access literatures directly supportive of the target literature and b) access literatures related to the target literature by some common or intermediary thread, will not provide

the insights and discoveries from these other disciplines that often result in innovations in the target discipline of primary interest (1).
3) Queries that are severely restricted in length, that rely in large measure on generic indexer-supplied terms, and have not been extensively iterated with the author-supplied language in the source database, will be inadequate in capturing the myriad ways in which different authors describe the same concept. They will also yield many records that are non-relevant to the main technical themes of the study.

In summary, these types of simple limited queries can result in two serious problems: a substantial amount of relevant literature is not retrieved, and a substantial amount of non-relevant literature is retrieved. As a result, the potential user is either overwhelmed with extraneous data, or is uninformed about existing valuable information, leading to potential duplication of effort and/ or R&D based on incomplete use of existing data. All the subsequent data processing, both human and computerized, cannot compensate for these deficiencies in the base data quality.

WHAT IS REQUIRED TO OBTAIN HIGH QUALITY INFORMATION RETRIEVAL AND ANALYSIS USING MODERN INFORMATION TECHNOLOGY?

The quality of a text mining study cannot exceed the quality of each of its components. For comprehensive access to the global S&T literature, and maximum extraction of useful information from this literature, four primary conditions are required.

1) A large fraction of the S&T conducted globally must be documented (INFORMATION COMPREHENSIVENESS).
2) The documentation describing each S&T project must have sufficient information content to satisfy the analysis requirements (INFORMATION QUALITY).
3) A large fraction of these documents must be retrieved for analysis (INFORMATION RETRIEVAL).
4) Techniques and protocols must exist for extracting useful information from the retrieved documents (INFORMATION EXTRACTION).

WHAT ARE THE ROADBLOCKS TO ACHIEVING HIGH QUALITY TEXT MINING/ INFORMATION RETRIEVAL OF THE GLOBAL S&T LITERATURE?

The approaches presently used by the majority of the technical community to address all four of these requirements have serious deficiencies.

1) *Information Comprehensiveness* is limited because there are many more disincentives than incentives for publishing S&T results (2), and therefore only a very modest fraction of S&T performed ever gets documented. Of the performed S&T that is documented, only a very modest fraction is included in the major databases. ***The contents of these knowledge repositories are determined by the database developers, not the S&T sponsors or the potential database users.*** Of the documented S&T in the major databases, only a very modest fraction is realistically accessible by the users. The databases are expensive to access, not very many people know of their existence, the interface formats are not standardized, and many of the search engines are not user-friendly.

2) *Information Quality* is limited because uniform guidelines do not exist for contents of the major text fields in database records (Abstracts, Titles, Keywords, Descriptors), and because of logic, clarity, and stylistic writing differences. The biomedical community has some advantage over the non-medical technical community in this area, since many medical journals require the use of Structured Abstracts (3). Compatibility among the contents of all record fields is not yet a requirement, and as our studies have shown (4), can lead to different perspectives of a technical topic depending on which record field is analyzed.

3) *Information Retrieval* is limited because time, cost, technical expertise, and substantial detailed technical analyses are required to retrieve the full scope of related records in a comprehensive and high relevance fraction process. Because much of the information technology community's focus is on selling search engine software, and automating the information retrieval process, they bypass the 'elbow grease' component required to get comprehensive and high signal-to-noise retrieval.

Our group has been developing information retrieval techniques using an iterative relevance feedback approach. The source database queries result in retrieval of very comprehensive source database records that encompass direct and supporting literatures with very high ratios of desired/ undesired records. Some of the queries consist of hundreds of terms (4), in stark contrast to the handful of phrases used in typical information retrieval. In

many cases, large queries are necessary to achieve the retrieval comprehensiveness and 'signal-to-noise' ratio required. Queries of a specific size are not a query development target of our group; rather, the query development process produces a query of sufficient magnitude to achieve the target objectives of comprehensiveness and high relevance ratio.

In each iterative step of our information retrieval process (5), a sample group of records retrieved with a previously modified query is classified into two categories: relevant to the central theme of interest, and non-relevant. Bibliometric and linguistic patterns of each category of the sampled records are examined, to generate terms that can be used to modify the query in order to increase relevant records (addition terms) and decrease non-relevant records (negation terms). The underlying assumption is that records in the source database that have the same linguistic patterns as the relevant records from the sample will have a high probability of being relevant, and records in the source database having the same linguistic patterns as the non-relevant records from the sample will also be non-relevant. Selection of such terms from the many thousands of candidate terms is a daunting task, and is extremely complex and time consuming.

To expand the relevant records retrieved, a phrase from the sample records should be added to the query if it:

1) appears predominately in the relevant record category;
2) has a high marginal utility (will retrieve a large ratio of relevant to non-relevant records) based on the sample;
3) has reasons for its appearance in the relevant records that are understood well; and
4) IS PROJECTED TO RETRIEVE ADDITIONAL RECORDS FROM THE SOURCE DATABASE (E.G., SCI) MAINLY RELEVANT TO THE SCOPE OF THE STUDY.

For multi-discipline source databases, application of these conditions can be complex. A recent example from the query development in a text mining study on the discipline of text mining illustrates this point. The phrase IR (an abbreviation for 'information retrieval' used in many SCI Abstracts) was characteristic of predominantly relevant sample records, had a very high absolute frequency of occurrence in the sample, and had a high marginal utility based on the sample. However, it was **not** 'projected to retrieve additional records from the source database mainly relevant to the scope of

the study'. A test query of IR in the SCI source database showed that it occurred in 65740 records dating back to 1973. Examination of only the first thirty of these records showed that IR is used in science and technology as an abbreviation for InfraRed (physics), Immuno-Reactivity (biology), Ischemia-Reperfusion (medicine), current(I) x resistance(R) (electronics), and Isovolume Relaxation (medical imaging). IR occurs as an abbreviation for information retrieval in probably one percent of the total records retrieved containing IR, or less. As a result, the phrase IR was not selected as a stand-alone query modification candidate.

Consider the implications of this real-world example. Assume a query consists of 200 terms. Assume 199 of these terms are selected correctly, according to the guidelines above. If the 200^{th} term were like IR above, then the query developer would have been swamped with an overwhelming deluge of unrelated records. ONE MISTAKE IN QUERY SELECTION JUDGEMENT can be fatal for a high signal-to-noise product.

Thus, the relation of the candidate query term to the objectives of the study, and to the contents and scope of the total records in the full source database (i.e., all the records in the SCI, not just those retrieved by the test query), must be considered in query term selection. The quality of this selection procedure will depend upon the expert(s)' understanding of both the scope of the study and the different possible meanings of the candidate query term across many different areas of R&D. *This strong dependence of the query term selection process on the overall study context and scope makes the 'automatic' query term selection processes reported in the published literature very suspect.*

A fully credible analysis requires expert domain knowledge on the part of the analyst(s). In addition, because any modern comprehensive information retrieval technique extracts information from many diverse literatures relative to a central problem of interest, experts having domain knowledge representing a diversity of backgrounds are required to exploit and interpret this multi-discipline information most efficiently.

4) *Information Extraction* is limited because the automated phrase extraction algorithms, required to convert the free text to phrases and frequencies of occurrence as a necessary first step in the text mining process, leave much to be desired. This is especially true for S&T free text, which the computer views as essentially a foreign language due to the extensive

use of technical jargon. Both a lexicon and technical experts from many diverse disciplines are required for credible information extraction. There is a widespread belief in the technical, and perhaps in the intelligence, communities that the combination of high speed computers with large memories, supported by intelligent agents and other artificial intelligence spin-offs, can produce automated/ semi-automated information extraction from large technical databases. This is not supported by experience, and continual combing of the technical literature shows no basis for this belief.

While the text mining processes and software structure the text information to aid the quality of the analysis, they do not bypass the requirement for in-depth rigor and thought characteristic of all serious technical investigations. As in the information retrieval step, detailed analysis of tens of thousands of phrases from many diverse technical literatures are required to integrate the data extracted into useful information. Again, in parallel with the information retrieval step, because the information technology community's focus is on selling information extraction software, and automating this information extraction process, they bypass the 'elbow grease' component required to get credible and useful results.

WHAT ARE SOME OF THE APPROACHES AVAILABLE TO OVERCOME THESE ROADBLOCKS?

1) Information Retrieval
S&T text mining can be used to enhance the retrieval of information from global S&T databases (4-7). Our group used all the S&T text mining components listed above to extract very comprehensive and highly relevant S&T information from global/ national semi-structured (free and structured text) S&T databases such as:

1a-Science Citation Index (compendium of 5600 journals addressing basic research)
1b-Engineering Compendex (compendium of over 5000 journals addressing applied research and technology development)
1c-MEDLINE (journal medical literature covering basic and applied research)
1d-National Technical Information Service (reports from U. S. government-sponsored basic research to advanced development)
1e-INSPEC (journal and conference proceedings covering basic research to technology development in physics, electronics, computing)

1f-RADIUS (narratives of U. S. government agency R&D programs)
1g-IBM and USPTO Patents (patent database)

2) Infrastructure Identification
S&T text mining can be used to identify the technology infrastructure (authors, journals, organizations) of a technical area (4-7). This infrastructure includes the authors (if known), journals, performing organizations, and countries. Such information is valuable for identifying experts for technical workshops and review panels, and for planning site evaluation visits. This information becomes critical for intelligence studies, where tracking of people and institutions, and analyzing time trends, is a central component of the analysis.

3) Literature-Based Discovery
S&T text mining can be used to discover new concepts or new relationships from literature, especially extrapolated from disparate literatures (1). Such information can be used to identify promising research or technology opportunities, and promising new directions for research. For intelligence applications, this approach can forecast the emergence of new unforeseen capabilities, based entirely on cutting-edge findings from global research laboratories.

This is an area of investigation that has completely fallen through the cracks. There is essentially one group that has published completed studies on different topics using this general technique (8-13), and these published studies have focused solely on the medical literature. A few other groups have published or presented concept papers (1, 14-18), using their own literature-based discovery techniques. Many more efforts are needed to test the applicability of competing techniques and the utility of different databases. Section 4 addresses the issue of literature-related discovery in far more detail.

4) Theme Identification
S&T text mining can be used for identifying the main technical themes or sub-themes in a large body of technical literature. Visual categorization of phrases allows technical taxonomies (classification schemes) to be generated (4-7). By categorizing phrases and counting frequencies, S&T text mining can also be used to estimate global levels of emphasis in technical areas or sub-areas. These results can be used as the basis for S&T adequacy or deficiency judgements (4).

5) Theme Relationship Identification

S&T text mining can be used to identify the relationships between technical themes, and between technical themes and infrastructure components (4-7). Much of the our present effort is focused on understanding and developing clustering approaches that will sharpen the groupings of common themes, and identify theme linkages from a variety of perspectives. We are presently using the more sophisticated clustering software in parallel with clustering technique development to link major technical themes, major technical themes with supporting technical themes, and technical themes with infrastructure components. Such linkages are important not only in the development of science and technology, but have important corporate and defense intelligence applications, and can provide direction for program and organizational restructuring based on technical content. Further, S&T text mining can generate taxonomies from the bottom-up, removing human subjectivity from the process to some extent.

6) Impact Roadmaps

S&T text mining can provide roadmaps of myriad research impacts (19, 20, 21). Such information is useful for impact tracking and subsequent S&T sponsor presentations. It provides performer organizations the ability to determine if the audience reached is the target audience. The Citation Mining approach we developed for these applications includes both bibliometric profiling and text mining. It is a sub-set of the more general trans-citation analysis, and has important consequences for Web-based corporate and national security intelligence information extraction.

CONCLUSIONS

The global S&T literature contains substantial technical data. For maximum benefit in conducting research and development, including the avoidance of duplication and the identification of promising opportunities, this data must be identified, retrieved, processed, and integrated. Much effort is being applied to information identification and retrieval, and sophisticated processes and tools exist presently. Little use is being made of these capabilities presently for the global S&T literature. This is especially serious for the medical community, since intense specialization in recent history has made many disparate but information-rich literatures unaccessible as new knowledge sources.

Far less effort is being devoted to the information processing and integration problems, especially the cutting-edge area of literature-based discovery and innovation. The technical community needs to be made aware of the potential of information technology for improving the conduct of S&T, the capabilities that exist presently, and the research required to deliver the full potential of information technology for this application.

REFERENCES FOR SECTION 2.

1. Kostoff, R. N. "Science and Technology Innovation". Technovation. 19, October 1999.
2. Kostoff, R. N., "The Underpublishing of Science and Technology Results", The Scientist, 1 May 2000.
3. Haynes, R.B., Mulrow, C. D., Huth, E. J., et al, "More informative abstracts revisited". Ann. Intern. Med. 1990; 113: 69-76
4. Kostoff, R. N., Green, K. A., Toothman, D. R., and Humenik, J. "Database Tomography Applied to an Aircraft Science and Technology Investment Strategy". Journal of Aircraft, 37:4, July-August 2000.
5. Kostoff, R. N., Braun, T., Schubert, A., Toothman, D. R., and Humenik, J. "Fullerene Roadmaps Using Bibliometrics and Database Tomography". Journal of Chemical Information and Computer Science. Jan-Feb 2000.
6. Kostoff, R. N., Eberhart, H. J., and Toothman, D. R. "Hypersonic and Supersonic Flow Roadmaps Using Bibliometrics and Database Tomography". Journal of the American Society for Information Science. 15 April 1999.
7. Kostoff, R. N., Eberhart, H. J., and Toothman, D. R., "Database Tomography for Technical Intelligence: A Roadmap of the Near-Earth Space Science and Technology Literature", Information Processing and Management,
8. Smalheiser, N.R., Swanson, D.R., "Assessing a Gap in the Biomedical Literature – Magnesium – Deficiency and Neurologic Disease", Neurosci Res Commun 15: (1), 1994.
9. Smalheiser, N.R., Swanson, D.R., "Calcium-Independent Phospholipase A (2) and Schizophrenia". Arch Gen Psychiat 55 (8), 1998
10. Smalheiser, N.R., Swanson, D.R., "Using ARROWSMITH: A Computer Assisted Approach to Formulating and Assessing Scientific Hypotheses" Comput Meth Prog Bio 57: (3), 1998.
11. Swanson, D.R., "Fish Oil, Raynauds Syndrome, and Undiscovered Public Knowledge", Perspect Biol Med .30: (1), 1986.

12. Swanson, D.R., Smalheiser, N.R., "An Interactive System for Finding Complementary Literatures: A Stimulus to Scientific Discovery". Artif Intell 91 (2), 1997.
13. Swanson, D.R., "Computer – Assisted Search for Novel Implicit Connections in Text Databases". Abstr Pap Am Chem S 217, 1999.
14. Gordon MD, Lindsay RK. Toward discovery support systems: A replication, re-examination, and extension of Swanson's work on literature-based discovery of a connection between Raynaud's and fish oil . Journal of the American Society for Information Science 47 (2): 116-128. Feb 1996.
15. Gordon MD, Dumais S. Using latent semantic indexing for literature based discovery. Journal of the American Society for Information Science 49 (8): 674-685. June 1998.
16. Weeber M, Klein H, de Jong-van den Berg LTW, et al. Using concepts in literature-based discovery: Simulating Swanson's Raynaud-fish oil and migraine-magnesium discoveries. Journal of the American Society for Information Science and Technology. 52 (7): 548-557. May 2001.
17. Stegmann J, Grohmann G. Hypothesis generation guided by co-word clustering. Scientometrics 56 (1): 111-135. 2003.
18. Hearst, M. A., "Untangling Text Data Mining", ***Proceedings of ACL 99, the 37th Annual Meeting of the Association for Computational Linguistics,*** University of Maryland, June 20-26, 1999.
19. Del Río, J. A., Kostoff, R. N., García, E. O., Ramírez, A. M., and Humenik, J. A., "Citation Mining Citing Population Profiling using Bibliometrics and Text Mining". Centro de Investigación en Energía, Universidad Nacional Autonoma de Mexico.
http://www.cie.unam.mx/W_Reportes.
20. Kostoff, R. N., and Del Rio, J. A., "The Impact of Physics Research", Physics World, June 2001.
21. Kostoff, R. N., Del Río, J. A., Humenik, J. A., García, E. O., Ramírez, A. M., "Citation Mining: Integrating Text Mining And Bibliometrics For Research User Profiling," JASIST. 52:13. 1148-1156 November 2001.

Section 3. Information Content in Medline Record Fields.

(based on Kostoff, R. N., Block, J. A., Stump, J. A., and Pfeil, K. M. "Information Content in Medline Record Fields". International Journal of Medical Informatics. 73:6. 515-527. June 2004)

OVERVIEW

Background: We have been conducting text mining analyses (extraction of useful information from text) of Medline records, using Abstracts as the main data source. For literature-based discovery, and other text mining applications as well, all records in a discipline need to be evaluated for determining prior art. Many Medline records do not contain Abstracts, but typically contain Titles and Mesh terms. Substitution of these fields for Abstracts in the non-Abstract records would restore the missing literature to some degree.

Objectives: Determine how well the information content of Title and Mesh fields approximates that of Abstracts in Medline records.

Approach: Select historical Medline records related to Raynaud's Phenomenon that contain Abstracts. Determine the information content in the Abstract fields through text mining. Then, determine the information content in the Title fields, the Mesh fields, and the combined Title-Mesh fields, and compare with the information content in the Abstracts.

Results: Four metrics were used to compare the information content related to Raynaud's Phenomenon in the different fields: total number of phrases; number of unique phrases; content of factors from factor analyses; content of clusters from multi-link clustering. The Abstract field contains almost an order of magnitude more phrases than the other fields, and slightly more than an order of magnitude more unique phrases than the other fields.

Each field used a factor matrix with fourteen factors, and the combination of all 56 factors for the four fields represented 27 separate, but not unique, themes. These themes could be placed in two major categories, with two sub-categories per major category: Auto-immunity (antibodies, inflammation) and circulation (peripheral vessel circulation, coronary vessel circulation). All four sub-categories included representation from each field. Thus, while the focus of the representation of each field in each sub-category

was moderately different, the four sub-category structure could be identified by analyzing the total factors in each field.

In the cluster comparison phase of the study, the phrases used to create the clusters were the most important phrases identified for each factor. Thus, the factor matrix served as a filter for words used for clustering. While clusters were generated for all four fields, the Title hierarchy tended to be fragmented due to sparsity of the co-occurrence matrix that underlies the clusters. Therefore, the Title clusters were examined at only the lower levels of aggregation.

The Abstract, Mesh, and Mesh + Title fields had the same first level taxonomy categories, autoimmunity and circulation. At the second level, the Abstract, Mesh, and Mesh + Title fields had the autoimmune diseases and antibodies sub-category in common. The Abstract and Mesh fields shared fascia inflammation as the other autoimmunity sub-category, while the other Mesh + Title sub-category focuses on vinyl chloride poisoning from industrial contact, and consequences of antineoplastic agents. However, in both cases, even though the words may be different, inflammation may be the common theme.

Conclusions: For taxonomy generation, especially at the higher levels, each of the four fields has a similar thematic structure. At very detailed levels, the Mesh and Title fields run out of phrases relative to the Abstract field. Therefore, selection of field (s) to be employed for taxonomy generation depends on the objectives of the study, particularly the level of categorization required for the taxonomy.

For information retrieval, or literature-based discovery, selection of the appropriate field again depends on the study objectives. If large queries, or large numbers of concepts or themes are desired, then the field with the largest number of technical phrases would be desirable. If queries or concepts represented by the more accepted popular terminology is adequate, then the smaller fields may be sufficient. Because of its established and controlled vocabulary, the Mesh field lags the Title or Abstract fields in currency. Thus, the Title or Abstract fields would retrieve records with the most explicitly stated current concepts, but the Mesh field would capture a

larger swath of fields that contained a concept of interest but perhaps had a wider range of specific terminology in the Abstract or Title text.

In addition, this section provides the first validated estimate of the disparity in information retrieved through text mining limited to Titles and Mesh terms relative to entire Abstracts. As much of the older biomedical literature was entered into electronic databases without associated Abstracts, literature-based discovery exercises that search the older medical literature may miss a substantial proportion of relevant information. On the basis of this section, it may be estimated that up to a log order more information may be retrieved when complete Abstracts are searched.

BACKGROUND

Text mining is the extraction of useful information from text (1-3). In its modern day incarnation, it refers to information technology-based extraction from very large volumes of text. One variant of text mining is literature-based discovery (4-8). It generates actual discovery from text only, using intermediate literatures that link the problem literatures (literatures that describe the problem to be solved) to discovery literatures (literatures that contain possible solutions to the problem).

The authors have been conducting a literature-based discovery study of Raynaud's Phenomenon. The first phase of this study is to compare the authors' results with those of Swanson's pioneering literature-based discovery paper on Raynaud's Phenomenon (6), and with subsequent papers that attempted to replicate Swanson's results (7-8).

Medline has been used as the source database for all of the above Raynaud's Phenomenon studies. The present study also used Medline Abstracts as the data source. However, in the 1975-1985 time frame that immediately preceded Swanson's initial Raynaud's Phenomenon publication, only 58% of the Medline records relating to Raynaud's Phenomenon contained Abstracts. Since a critical element of discovery requires establishing that prior art does not exist, all prior records need to be evaluated. Some type of information from these non-Abstract records needs to be included in the empty Abstract field. The only other type of text-based information available from Medline is contained in the Title and Mesh fields. This leads to the question: how well does the information content in the Title and Mesh fields approximate that in the Abstract fields?

Most of the studies comparing text fields in a full text document or database record have the objective of comparing information retrieval. In examining different approaches to indexing medical literature, Hersh and Hickam (9) found that text word indexing is more effective than MESH term indexing. In a test of whether Abstract words occurred as frequently as could be expected from full text statistical analysis, Su et al (10) found that 96-98% of the Abstracts tested were not significantly different from random samples of the articles they represented. However, from a retrieval perspective, Johnson et al (11) found that searching the full text version of NEJM retrieved a larger number of records than Medline. This was due primarily to methodology terms found in the text but not in the Title or Abstract. Additionally, in Medline, two indexer-supplied fields (descriptors and publication types) retrieved 11-89% more than Title or Abstract alone. This important role played by MESH in enhancing information retrieval was supported by Srinivasan (12). However, while MESH terms can enhance information retrieval, their limitations include the Indexer Effect (13), indexer consistency in Medline (14), indexer-searcher vocabulary mis-matches (15), and lag in adopting new terminology. Even concept hierarchies to aid the management of controlled vocabularies, such as the Metathesaurus of the Unified Medical Language System (16-17), may not always incorporate evolving topics in a timely manner (18).

Mijnhout et al (19) concluded that, for comprehensive retrieval, both MESH terms and text words should be used in a search strategy, implying a disparity between the fields. He also concluded that, for comprehensiveness, multiple databases should be searched. None of these articles have focused on information content exclusively, and used the combined clustering and unique phrase occurrence approach that will be described in this paper.

OBJECTIVES

This study will determine how well the information content in the Title and Mesh fields of Medline records approximates that contained in the Abstract field.

APPROACH

General

The approach consists of two components: quantitative and qualitative. The quantitative component compares total and unique words in each field. The qualitative component compares category taxonomies in each field. The former provides detail, while the latter provides structure.

Specific

The query "RAYNAUD'S DISEASE OR RAYNAUD*[TW]" (restricted to 1975-1985) was inserted into the PubMed Search engine for Medline, and retrieved 932 records with Abstracts (these are the 58% of the total records in the 1975-1985 time frame that contained Abstracts). The contents of the Abstracts field, Titles field, Mesh field, and Mesh-Titles field were placed in separate databases. The information content in each of these four databases (Abstract, Title, Mesh, and combined Title-Mesh) was compared using text mining. Both qualitative and quantitative metrics were used for the comparison.

As with any quantitative comparison procedure, a critical item is the selection of appropriate metrics. In comparing any two databases for this study, the main focus is on assessing the importance of unique phrases and phrase patterns in each database relative to the other databases. This assessment is made at two levels of aggregation.

At the lower phrase-focused level, the numbers of total and unique phrases, and the significance of each phrase, are compared among the databases. Because of the present study's context of information for literature-based discovery, and literature-based discovery's extensive use of word/ phrase matching among documents from different retrievals, numbers of phrases becomes significant in the matching process. The greater the number of phrases, the larger the dimensions describing each concept, and the greater the probability that two equivalent concepts will exhibit some overlap of their component phrases. At the higher phrase-pattern focused level, the phrase clusters and overall taxonomies are compared among the databases.

1) Lower Level Comparison

The number of phrases is relatively straight-forward, although not quite as simple as it would appear superficially. It is easy to count phrases produced

by a Natural Language Processor, or other type of phrase generator. It is more complex to count the subset of generated phrases that have high technical content, and that would be used in the performance of an actual text mining study. Judgement is required to translate the raw phrase generator outputs to useful phrases.

Even the metrics for classifying a phrase as "high technical content" are ill-defined. High technical content is a function of context, and classifying a phrase in isolation from its context is fraught with error. In the phrase clustering process presented later in this paper, a method called factor matrix filtering is used to insure that high technical content phrases only are used for clustering. Only those phrases that have substantial influence on determining the themes of the factors in a factor analysis are used for the clustering, and these phrases become high technical content by virtue of their function and context.

The significance of each phrase is far more complex to obtain, since it depends on the context of the analysis to be performed. If a macro level analysis is the objective, such as development of a higher level taxonomy, then a database that is deficient in a few phrases relative to some other more detailed database may affect the final taxonomy relatively little (hypothetical at this point). However, if a micro level analysis is the objective, such as literature-based discovery (4-8), then a deficiency of a very few phrases may be crucial to the results, if the discovery elements are contained within these missing phrases.

The metrics used to compare the different databases for their impact on taxonomy generation reflect the above issues. These metrics focus on counting the differences in unique phrases contained in each database. The metrics do not address the significance of the absent phrases, since that requires judgement about the quality of the application. The significance of the phrase absences from any database relative to other databases is judged qualitatively.

2) Higher Level Comparison

The high technical content phrases in each database are aggregated into clusters, and the clusters are integrated to form a taxonomy. The taxonomies

are then compared for structural and conceptual differences. Two statistical approaches for taxonomy generation are used, factor matrix and multi-link clustering.

Since each phrase, phrase cluster, and taxonomy addresses some aspect of Raynaud's Phenomenon, an overview of Raynaud's Phenomenon will be presented before discussing the results. Because the main Raynaud's terminology used in the literature is not consistent (in many cases, Raynaud's Phenomenon is used interchangeably with Raynaud's Disease or Raynaud's Syndrome), the overview will include the distinction among these Raynaud variants.

Raynaud's Phenomenon Overview

Raynaud's Phenomenon is a condition in which small muscular arteries and arterioles, most commonly in the fingers and toes, go into spasm (contract) and cause the skin to turn pale (blanching) or a patchy red (rubor) to blue (cyanosis). While this sequence is normally precipitated by exposure to cold, and resolves with subsequent re-warming, it can also be induced by anxiety or stress. Blanching represents the ischemic (lack of adequate blood flow) phase, caused by digital artery vasospasm. Cyanosis results from de-oxygenated blood in capillaries and venules (small veins). Upon re-warming, a hyperemic phase ensues, causing the digits to appear red.

Raynaud's Phenomenon can be a primary or secondary disorder. When Raynaud's symptoms appear alone without an apparent underlying medical condition, it is referred to as Primary Raynaud's Phenomenon or, formerly, as Raynaud's Disease. In this condition, the blood vessels appear anatomically normal after the ischemic events. When an identifiable cause or a specific associated disease accompanies Raynaud's symptoms, it is referred to as Secondary Raynaud's Phenomenon. The auto-immune disorders, or conditions in which a person produces an immune response against his or her own tissues, are the typical medical conditions associated with Secondary Raynaud's. In these cases, Raynaud's Phenomenon may be more serious than in Primary Raynaud's and may result in blood vessel scarring and long-term consequences. When Raynaud's Phenomenon is associated with occupational activities, such as vibrating machinery or repetitive activity, it is often referred to as Occupational Raynaud's. Similarly, Secondary Raynaud's may be precipitated by exposure to harmful

chemical compounds such as vinyl chloride, or to toxic therapeutic agents such as certain cancer chemotherapy drugs.

Thus, while the symptoms and signs of Raynaud's Phenomenon occur as a direct consequence of reduced blood flow due to reversible blood vessel constriction, the underlying etiology may be a function of several parameters that affect blood flow. These include:

*Inflammation from the auto-immune disorders that can cause swelling and thereby constrict blood vessels;
*Increased sympathetic nervous system activity, that can affect the timing and duration of the blood vessel muscular contractions that cause constriction;
*Heightened digital vascular reactivity to vaso-constrictive stimuli, that causes the blood vessels to over-react and over-contract;
*Deposits along the blood vessel walls that can reduce blood flow and increase the flow sensitivity to contraction stimuli;
*Blood rheological properties that offer additional resistance to blood flow, and magnify the impact of blood vessel constriction;
*Blood constituents and hormones that can act as vaso-constrictors or vaso-dilators.

RESULTS

Phrase frequencies were generated for ten years of Mesh Terms, Titles, Mesh-Titles, and Abstracts from Medline records focused on Raynaud's Phenomenon (1975-1985). A sampling across frequency bands in the larger text fields showed that about 2/3 of the phrases could be classified as high technical content.

1) Quantitative and Qualitative Properties of Phrases in each Field

Table 1A lists the number of phrases in each field, incorporating single, double, and triple word phrases. The first column represents the databases in which the phrases appeared. The Abstracts contain almost an order of magnitude more phrases than the other two fields. Also, the low number of separate Mesh phrases reflects the restrictions imposed by a controlled

vocabulary: less diversity, more uniformity. About 90% of the Abstract phrases tend to be unique, whereas slightly less than half of the Title and Mesh phrases are unique.

TABLE 1A – NUMBER OF PHRASES IN EACH FIELD

FIELD	TOTAL	UNIQUE
Abstract	44,029	40114
Title	5941	2780
Mesh	2735	1237
Mesh + Title	7950	

TABLE 1B – 'UNIQUE' TITLE PHRASES AND ABSTRACT REPRESENTATIONS

TITLE	ABSTRACT
ANGIOLOGIC	ANGIOLOGICAL
HAEMORHEOLOGY	HAEMORHEOLOGICAL
FINGER ULCER	FINGER ULCERATIONS
DISEASE GROUPS	DISEASE GROUP
ARTERIAL INFUSION	INTRA-ARTERIAL (IA) INFUSION
ANTIBODIES AGAINST SCL 70	ANTIBODIES AGAINST SCL-70
PERIPHERAL CIRCULATORY DISORDERS	PERIPHERAL CIRCULATION DISORDER

At this point, some readers may question the entries in Table 1A that show unique Title phrases relative to the Abstracts. Wouldn't every important technical word/ phrase in the Title appear in the Abstract? The answer depends on the definition of unique. If 'unique' is defined as 'identical', the answer is no. If 'unique' is defined as 'very similar', the answer is yes. For example, Table 1B contains some of the 'unique' Title phrases (as defined in Table 1A), and their Abstract representations.

Table 2 lists the 20 highest frequency high technical content phrases for each of the four databases (fields). All four databases share the following phrases: Raynaud, Disease, Systemic, Blood, Scleroderma, Antibodies, Finger(s), Skin, Lupus Erythematosus, Connective Tissue, and Vibration. The phrases are mainly single word, with some double word included. The high frequency phrases are relatively simple and generic, and major differences among the fields are not evident at this high frequency level of description.

TABLE 2 – HIGHEST FREQUENCY NON-DUPLICATIVE HIGH TECHNICAL CONTENT PHRASES

ABSTRACT PHRASES	MESH PHRASES	TITLE PHRASES	TITLE + MESH PHRASES
PATIENTS	HUMAN	RAYNAUD	RAYNAUD
RAYNAUD	DISEASE	DISEASE	DISEASE
DISEASE	RAYNAUD	PHENOMENON	HUMAN
PHENOMENON	BLOOD	SYNDROME	BLOOD
SYNDROME	COMPLICATIONS	SYSTEMIC	COMPLICATIONS
SYSTEMIC	THERAPY	PATIENTS	DRUG
BLOOD	DIAGNOSIS	SCLERODERMA	THERAPY
SCLERODERMA	IMMUNOLOGY	TREATMENT	DIAGNOSIS
TREATMENT	ADULT	SCLEROSIS	IMMUNOLOGY
ANTIBODIES	ETIOLOGY	TISSUE	ETIOLOGY
SKIN	MIDDLE	CONNECTIVE	DISEASES
TISSUE	DISEASES	LUPUS	MIDDLE
COLD	PATHOLOGY	BLOOD	SYSTEMIC
SYMPTOMS	PHYSIOPATHOLOGY	ERYTHEMATOSUS	PATHOLOGY
VASCULAR	THERAPEUTIC	ANTIBODIES	THERAPEUTIC
FLOW	SYSTEMIC	VIBRATION	SCLERODERMA
FINGER	SUPPLY	VASCULAR	SUPPLY
SCLEROSIS	SCLERODERMA	PERIPHERAL	SYNDROME
CONNECTIVE	SUPPORT	PRIMARY	SUPPORT
TEMPERATURE	RADIOGRAPHY	DISEASES	ADVERSE

Table 3 lists single word phrases in each database that are unique; i.e., not contained in any of the other databases. The Title-Abstract caveats with respect to uniqueness apply here as well.

TABLE 3 – UNIQUE PHRASES IN EACH DATABASE

	ABSTRACT	TITLES	MESH
1	ACETYLSALICYLIC	ADENINE	ABORTION
2	ATHEROSCLEROSIS	CRYOABLATION	BIOMECHANICS
3	ALANINE	EFAMOL	CORTEX
4	ANOREXIA	MALEATE	CRYOSURGERY
5	ANTIGLOBULIN	PREDILECTION	DYSPEPSIA
6	APERISTALSIS	WEARDALE	EMBRYOLOGY
7	CORTISOLAEMIA	SULPIRIDE	HYPOTHERMIA
8	ENDANGIITIS	NIACIN	ACETYLGLUCOSAMINIDASE
9	FLUORESCEIN-CONJUGATED	KAPOSI	ADENOSINE
10	HEXOPAL	FUROSEMIDE	ADULTORUM

The Mesh phrases, on average, appear somewhat more generic than the Title or Abstract phrases. They represent a controlled vocabulary, and can contain additional information to that present in Title or Abstract fields (and vice versa).

Table 4 contains phrases shared by all fields. These are the standard well-known phrases associated with Raynaud's, especially the more focused multi-word phrases.

TABLE 4 – SHARED PHRASES AMONG ALL FIELDS

	SINGLE WORDS	DOUBLE WORDS
1	ABDOMINAL	ACANTHOSIS NIGRICANS
2	TOXOPLASMOSIS	CONNECTIVE TISSUE
3	ACANTHOSIS	FACIAL HEMIATROPHY
4	CORPUSCLES	MULTIPLE SCLEROSIS
5	FIBROSIS	PLASMA EXCHANGE
6	GUANETHIDINE	RESPIRATORY FUNCTION
7	LEUKOCYTES	SCLERODERMA SYSTEMIC
8	METOPROLOL	SKIN MANIFESTATIONS
9	ORTHOSTATIC	THROMBOANGIITIS OBLITERANS
10	SUBCUTANEOUS	VASCULAR DISEASES

2) Taxonomies for all Four Databases

The previous section has shown differences among the databases at the micro, or individual phrase, level. Both the differences in total number of phrases, and number of unique phrases, were shown. For a crude measure of significance of these phrase differences, some examples of unique phrases were also shown.

However, differences at the macro, or aggregated phrase, level were not shown. These aggregated phrases, or clusters, can be thought of as concepts. It would be useful to ascertain whether there are conceptual differences among the databases.

There are two measures of importance in measuring conceptual differences. One measure is difference in structure of the overall database taxonomy, i.e., are there any concepts missing from any of the databases, and even if not, do all the concepts bear the same relationships across databases? The other measure is differences in resolution of each concept, i.e., how does the information content in each cluster vary across the different databases?

This section addresses the taxonomy structure metric. Two approaches are used to develop taxonomies for each database. A factor matrix is used first, followed by a multi-link clustering method. The combination has been used in previous text mining studies (e.g., 20), although the present application uses a synergistic combination of the two methods that offers substantial improvement in the quality of the resultant clusters. The factor matrix, which imports a relatively large number of words, is used as a filter for the words subsequently input to the clustering algorithms. This results in a selection of context-dependent words for input to the clustering algorithm, and produces relatively well defined clusters compared to the context-independent methods. In addition, the present application uses single words for clustering, rather than the multi-word phrases of previous applications. While some of the technical detail is lost by excluding the ordering information contained in multi-word phrases, inclusion of all single words compensates for the elimination of multi-word phrases due to the selection algorithm of the Natural Language Processor.

2A. Factor Matrix Filtering

Factor analysis aims to reduce the number of variables in a system, and detect structure in the relationships among variables. One of the key challenges in factor analysis is defining the number of factors to select. Different approaches have been suggested in the literature, but the two most widely used are the Kaiser criterion (21-22), and the Scree test (23). The Kaiser criterion states that only factors with eigenvalues greater than unity should be retained, essentially requiring that a factor extracts at least as much variance as the equivalent of one original variable. The Scree test plots factor eigenvalue (variance) vs factor number, and recommends that only those factors that extract substantive variance be retained. Operationally, the factor selection termination point becomes the ‘elbow’ of the plot, the point where the slope changes from large to small.

In most previous studies performed by the first author, the Kaiser criterion has been used to select the number of factors for the factor matrix. These previous studies have used an Excel add-in to generate the factor matrices and, due to Excel's limitations on columns, have been limited approximately to 250 x 250 correlation matrices, or 250 words. The Kaiser criterion has yielded factor numbers in the range of 20-45, considered a reasonable number for analysis. However, in the present validation study, another software package that did not require Excel was used (TechOasis), and many more words were used for the correlation matrix. The Kaiser criterion yielded hundreds of factors, a number far too large for detailed factor analysis, and of questionable utility, since many of the eigenvalues were not too different from unity. The Scree Plot was examined, and used to select the number of factors for analysis.

Once the desired value of the Scree Plot 'elbow' has been determined, and the appropriate factor matrix has been generated, the factor matrix can then be used as a filter to identify the significant technical words for further analysis. Specifically, the factor matrix can complement a basic trivial word list (e.g., a list containing words that are trivial in almost all contexts, such as 'a', 'the', 'of', 'and', 'or', etc) to select context-dependent high technical content words for input to a clustering algorithm. The factor matrix pre-filtering will improve the cohesiveness of clustering by eliminating those words that are trivial words operationally in the application context.

In the factor matrix used, the rows are the words and the columns are the factors. The matrix elements Mij are the factor loadings, or the contribution of word i to the theme of factor j. The theme is determined by those words that have the largest absolute values of factor loading. Each factor had a positive value tail and negative value tail. For each factor, most of the time, one of the tails dominated in terms of absolute value magnitude. This dominant tail was used to determine the central theme of each factor.

For the first step in the factor matrix filtering process, the factor loadings in the factor matrix were converted to absolute values. Then, a simple algorithm was used to automatically extract those high factor loading words at the dominant tail of each factor. If word variants were on this list (e.g., singles and plurals), and their factor loadings were reasonably close (24), they were conflated (e.g., 'agent' and 'agents' were conflated into 'agents', and their frequencies were added). A few words were eliminated manually,

based on factor loading and estimate of technical content. Basically, any word that did not have a high absolute value of factor loading for at least one factor was eliminated from the subsequent clustering.

Before the clustering is described, the factors that formed the basis of the factor matrix filtering for all four fields will be described. Table 5 lists the number of words that were input to the factor matrix algorithm for each field. In each case, addition of succeeding lower frequency words (from the raw data word list ordered by frequency) made the factor matrix difficult to generate.

TABLE 5 – NUMBER OF WORDS INPUT TO FACTOR MATRIX

ABSTRACT	TITLE	MESH	M+T
659	428	465	519

Table 6 lists the different factor themes that can be extracted from all four fields, and identifies the factor(s) in which the theme occurs for each field. A more expanded description of the non-redundant themes is presented in Appendix 1.

TABLE 6 – FACTOR THEMES ASSOCIATED WITH EACH RECORD FIELD

FACTOR THEME	RECORD FIELD			
	ABST	TITLE	MESH	M+T
ANTIBODIES AND AUTOIMMUNE DISEASES	1	6	1	1
RAYNAUD'S SYNDROME-RELATED AUTOIMMUNE DISEASES	14		8	6
SYSTEMIC LUPUS ERYTHEMATOSUS CLASSIFICATION		5		
ADRENAL CORTEX HORMONES FOR LUPUS			5	5
SCLEROTIC AUTOIMMUNE DISEASES	6		13	
ENDOTHELIAL CELL ACTIVITY IN PSS		10		
MIXED CONNECTIVE TISSUE DISEASE		9		
CIRCULATING IMMUNE COMPLEXES	12	8	3	11
DEFICIENCIES OF COMPLEMENT COMPONENTS		13		
FASCIA-FOCUSED INFLAMMATION	13		1	13
CORONARY CIRCULATION AND HYPERTENSION	4		10	
SMOOTH MUSCLE RESPONSE TO DRUGS			14	14
PLASMA LIPID FRACTION CONTROL			11	3
CARDIAC BLOOD SUPPLY OBSTRUCTIONS		11		
DOUBLE-BLIND VASODILATOR TRIALS	2	4	2	6, 9
CALCIUM CHANNEL BLOCKERS			9	12

BIOFEEDBACK TRAINING FOR IMPROVED CIRCULATION	10		6	10
REDUCED PLATELET AGGREGATION VASODILATORS	5	12	2, 8	2, 8
PERIPHERAL CIRCULATORY SYSTEM VASODILATION	11			
FINGER BLOOD FLOW MEASUREMENTS	8	3		
NAIL-FOLD CAPILLARY MICROSCOPY	9	2		11
TRANSCUTANEOUS NERVE STIMULATION/VASODILATION		1		14
SURGICAL/ NERVE BLOCK CIRCULATION OBSTRUCTION SOLUTIONS	7		7, 12	7
CARPAL TUNNEL SYNDROME		14		
VIBRATION-BASED CIRCULATION PROBLEMS	3	7	13	9, 13
VINYL CHLORIDE EFFECTS ON BLOOD FLOW AND PROPERTIES	14	7	7, 12	4
TREATMENTS FOR CHEMICALLY-INDUCED NEOPLASMS			4	

At the highest level, the themes can be divided into two categories, auto-immunity and circulation. This division into these two categories is suggested by the data in the above table, is shown more sharply by the multi-link hierarchical structures, and reflects the experience of medical practice for Raynaud's Phenomenon. The first ten themes fit into the auto-immunity category, and the remainder fit into the circulation category.

At the next hierarchical level, the auto-immunity category can be divided into antibody and inflammation sub-categories, but other divisions are possible as well. The circulation category can be sub-divided into coronary vessel circulation and peripheral vessel circulation.

All fields are represented in some themes of each of the four second-level categories, although all fields are not represented in all themes. This should not be interpreted that a field-theme combination that is absent from the matrix means that the theme is absent from the field. It only means that, within the resolution afforded by a fourteen factor matrix, a specific theme was not among the fourteen major themes for that field.

For example, FASCIA-FOCUSED INFLAMMATION is not listed as a Title theme on the matrix. However, EOSINOPHILIC FASCIITIS appears in the Title words used to form the factor matrix. Also, BIOFEEDBACK TRAINING FOR IMPROVED CIRCULATION is not listed as a Title theme in the matrix, but BIOFEEDBACK appears in the Title word list. To fill in the blanks on Table 6, matrices with more factors would have to be generated, in turn producing a larger version of Table 6 with additional blanks.

Because of the smaller number of Title words, the themes in each category in which the Title is represented tend to be more specific relative to those of the Mesh or Abstract. The larger numbers of words in the Abstract or Mesh fields allow more generalized themes to populate the categories.

2B. Multi-Link Hierarchical Clustering

The filtered and conflated words resulting from the factor matrix filtering were input to the WINSTAT clustering algorithm, and the Average Link option was selected for clustering. The major structures, or clusters, from the resulting dendrogram (a tree-like structure that shows how the individual words cluster into groups in a hierarchical structure) were analyzed, and compared for the four fields. In most cases, only the top three hierarchical levels of each field's taxonomy were described.

First Taxonomy Level

The Abstract, Mesh, and Mesh + Title fields have the same first level taxonomy categories, auto-immunity and circulation, and these two categories are sharply delineated for all three fields.

The hierarchical structure of the Titles dendrogram is very different from that of the Abstracts dendrogram. When the Abstracts clusters divide at each lower level in the hierarchy, they split into sub-clusters that are of reasonably similar magnitude (usually, not always) and usually have complementary themes. When the Titles clusters divide, they are splitting into sub-clusters of very different magnitudes with themes that are not very complementary.

For example, the top level split of the Abstracts dendrogram is into two clusters containing 90 and 162 words. The themes of these clusters are auto-immunity and circulation, respectively, the two main complementary themes of Raynaud's Phenomenon. The top level split of the Titles dendrogram is into two clusters containing 4 and 249 words. The themes of these clusters are hereditary deficiency for the smaller cluster, and auto-immunity and circulation for the larger cluster. The next level split for the circulation cluster in the Abstracts dendrogram is into two clusters containing 25 and

137 words. The themes of these clusters are coronary circulation and peripheral circulation, respectively, the two main complementary themes of circulation. The next level split for the auto-immunity and circulation cluster in the Titles dendrogram is into two clusters containing 25 and 224 words. The themes of these clusters are controlled double blind trials of coronary vasodilators and antihypertensives, and autoimmunity and peripheral circulation and other aspects of coronary circulation, respectively. Thus, rather than splitting high level clusters into the next level clusters in a hierarchically conceptual process, the Titles dendrogram is stripping out a low level very detailed cluster from a high level cluster division. This type of structure has been found in studies of applications of a particular research or technology discipline, where the diversity of the applications produces numerous relatively unrelated themes. For the Titles, the relatively low word frequencies produce a sparse co-occurrence matrix, and the resultant clusters appear fragmented. The highest level categorizations in the Title field are not viewed as meaningful, and will not be addressed until the third level clusters are discussed.

Second Taxonomy Level

Table 7 shows all the second level categories from the Abstract, Mesh, and Mesh + Title fields, and the representation of each field in each category. Three asterisks denote fully applicable, while two asterisks denote a partial match.

TABLE 7 – SECOND LEVEL CATEGORY THEMES IN EACH FIELD

CATEGORY	ABST	MESH	M+T
AUTOIMMUNE DISEASES/ ANTIBODIES	***	***	***
INFLAMMATION/ FASCIA	***	***	
PERIPHERAL VESSEL CIRCULATION	***	**	
CORONARY VESSEL CIRCULATION	***	**	
PERIPHERAL AND CORONARY VESSEL CIRCULATION	**	***	
CHEMICALLY/ CHEMOTHERAPEUTICALLY-INDUCED DISEASES		***	***
NON-PSYCHOLOGICAL CIRCULATION TREATMENTS			***
PSYCHOLOGICAL CIRCULATION TREATMENTS			***

In all three fields, ‘auto-immune diseases and antibodies’ is the only common category. Interestingly, it was the first factor in the three factor matrices as well. The main difference between the Mesh and Abstract fields at this second hierarchical level is that the Mesh gives more recognition to consequences of vinyl chloride poisoning from industrial contact, and consequences of anti-neoplastic agents. Addition of the Title field to the Mesh field has the further effect of providing more recognition to psychologically-based treatments for improving circulation (biofeedback and autogenic training).

Lowest Level-Elemental Clusters

Table 8 shows all the third level categories and sub-themes from the Abstract, Title, Mesh, and Mesh + Title fields, and the representation of each field in each category. Three asterisks denote fully applicable, while two asterisks denote a partial match.

TABLE 8 – THIRD LEVEL CATEGORY THEMES IN EACH FIELD

THIRD LEVEL CLUSTER THEMES AND SUB-THEMES

CATEGORY	ABST	TITLE	MESH	M+T
ANTIBODIES	***	***	***	***
SCLEROTIC AUTOIMMUNE DISEASES	***	***	***	***
RAYNAUD'S SYNDROME-RELATED AUTOIMMUNE DISEASES	***	**	***	***
CREST SYNDROME AUTOIMMUNE DISEASES	***	***	***	***
CIRCULATING IMMUNE COMPLEXES	***	**	***	***
LIVER ABNORMALITIES IN RHEUMATIC DISEASES	**	***	***	***
FASCIA-RELATED INFLAMMATION	***	***	***	***
DOUBLE-BLIND CLINICAL TRIALS FOR VASODILATORS	***	***	***	***
VASODILATORS FOR REDUCED PLATELET AGGREGATION	***	***	***	***
TRANSCUTANEOUS NERVE STIMULATION		***		***
ARTERIAL OCCLUSIONS IN EXTREMITIES	**	***	***	***
FINGER BLOOD FLOW AND TEMPERATURE MEASUREMENTS	***	**	***	***
VIBRATION EFFECTS ON NERVOUS SYSTEM AND CIRCULATION	***	***	***	***
VINYL CHLORIDE EFFECTS ON NERVOUS SYSTEM AND CIRCULATION	***	***	***	***
CHEMOTHERAPY TOXICITY	**	***	***	***
NAILFOLD CAPILLARY MICROSCOPY	***	***		
CARDIOVASCULAR/ PULMONARY CIRCULATION PROBLEMS	***	**	***	***
BIOFEEDBACK AND AUTOGENIC TRAINING	***	*	***	***

Most of the sub-themes are covered in all fields, although there are some notable absences, and they tend to occur in the circulation category. TRANSCUTANEOUS NERVE STIMULATION is identified in Title and Mesh + Title, but not in Abstract or Mesh, while NAILFOLD CAPILLARY MICROSCOPY is specifically identified in Abstract and Title, but not in Mesh or Mesh + Title.

DISCUSSION AND CONCLUSIONS

Four fields (Abstract, Title, Mesh, Mesh + Title) in 932 Medline records were compared for information content. Four metrics were used to assess information content: 1) total number of phrases; 2) number of unique phrases; 3) factors; 4) clusters.

The Abstract field contains almost an order of magnitude more phrases than the other fields, with the difference becoming more exascerbated as the words per phrase increases. Even though the Mesh field tends to be larger than the Title field in volume, it has about half the total number of phrases. This is due to the controlled vocabulary limiting the phrase diversity relative to the unrestricted vocabulary of the Titles.

The difference in number of unique phrases between the Abstract field and the Title or Mesh fields is even more pronounced. These large differences are not evident when the high frequency phrases in each field are compared. At this end of the spectrum, many of the phrases tend to be the more generic representations of Raynaud's Phenomenon, and many are shared in common. The phrase differences tend to be more evident and pronounced at the lower frequency end of the spectrum.

In the theme comparison phase of the study, fourteen factors were used based on Scree plots and standardization. The Abstract field allowed the largest number of words to generate the factor matrices, while the Title field allowed the smallest number of words. These numbers differed by about 50%.

About 27 separate, but not unique, themes could be discerned from all the factors combined. These themes could be placed in two major categories, with two sub-categories per major category: Auto-immunity (antibodies, inflammation) and circulation (peripheral vessel circulation, coronary vessel circulation). All four sub-categories included representation from each field. Thus, while the focus of the representation of each field in each sub-category was moderately different, the four sub-category structure could be identified by analyzing the total factors in each field.

In the cluster comparison phase of the study, the phrases used to create the clusters were the most important phrases identified for each factor. Thus, the factor matrix served as a filter for words used for clustering. All clusters were based on about 250 words. A hierarchical multi-link aggregation clustering technique was used to form the clusters.

While clusters were generated for all four fields, the Title hierarchy tended to be fragmented due to sparsity of the co-occurrence matrix that underlies the clusters. Therefore, the Title clusters were examined at only the lower levels of aggregation.

The Abstract, Mesh, and Mesh + Title fields had the same first level taxonomy categories, auto-immunity and circulation. At the second level, the Abstract, Mesh, and Mesh + Title fields had the auto-immune diseases and antibodies sub-category in common. The Abstract and Mesh fields shared fascia inflammation as the other auto-immunity sub-category, while the other Mesh + Title sub-category focuses on vinyl chloride poisoning from industrial contact, and consequences of antineoplastic agents.

This latter difference illuminates previous observations that the different record fields sometimes operate at different meta-levels of description. Many antineoplastic agents can produce specific organ inflammation, myocutaneous inflammation, or systemic inflammation during the course of treatment. This effect can be referenced as inflammation, antineoplastic agents, or both. While the theme can be different superficially in the different fields, as in the present case, underneath the theme/ concept may be the same, but expression occurred at different meta-levels.

At the next level of categorization, most of the sub-categories are covered in all fields. There are some notable absences, and they tend to occur in the circulation category. For example, TRANSCUTANEOUS NERVE STIMULATION is identified in Title and Mesh + Title, but not in Abstract or Mesh, while NAILFOLD CAPILLARY MICROSCOPY is specifically identified in Abstract and Title, but not in Mesh or Mesh + Title.

Finally, what is the importance of the differences among the fields summarized in the present study? All levels of text mining, ranging from standard information retrieval to the more exotic literature-based discovery, tend to access records through phrase matching. As the results imply, there could be substantial differences in numbers and types of records retrieved, depending on which fields are accessed by the search engines.

For high level taxonomy generation, the field differences are less severe, but when lower level taxonomic detail is required, then the differences become important. The Title and Mesh fields have limited numbers of different phrases at the lower frequencies relative to the Abstracts, and this translates in differences in diversity of detail possible. For literature-based discovery in particular, access to related and disparate literatures will be limited due to sheer phrase volume and diversity limitations. The predominant publishing group in literature-based discovery (4, 6) has used Title and Keyword phrases for information processing almost exclusively. Use of Abstracts should result in much more literature content accessed.

REFERENCES FOR SECTION 3.

1. Hearst, M. A. Untangling Text Data Mining. Proceedings of ACL 99, the 37th Annual Meeting of the Association for Computational Linguistics. University of Maryland. June 20-26, 1999.
2. Trybula, W.J. Text Mining. Annual Review of Information Science and Technology. 34. 385-419. 1999
3. Losiewicz, P., Oard, D., and Kostoff, R. N. Textual Data Mining to Support Science and Technology Management. Journal of Intelligent Information Systems. 15. 99-119. 2000.

4. Swanson, D.R, and Smalheiser, N.R. Implicit Text Linkages Between Medline Records: Using Arrowsmith as an Aid to Scientific Discovery. Library Trends. 48 (1): 48-59. 1999.
5. Kostoff, R. N. Science and Technology Innovation. Technovation. 19:10. 593-604. 1999.
6. Swanson, D.R. Fish Oil, Raynauds Syndrome, and Undiscovered Public Knowledge. Perspectives in Biology and Medicine. 30 (1): 7-18. 1986.
7. Gordon, M.D., Lindsay, R.K. Toward Discovery Support Systems: A Replication, Re-Examination, and Extension of Swanson's Work On Literature-Based Discovery of a Connection Between Raynaud's And Fish Oil. Journal of the American Society for Information Science. 47 (2): 116-128. 1996
8. Weeber, M, Klein, H, de Jong-van den Berg L..W., Vos, R. Using Concepts in Literature-Based Discovery: Simulating Swanson's Raynaud-Fish Oil and Migraine-Magnesium Discoveries. Journal of the American SocietyfFor Information Science and Technology. 52 (7): 548-557. 2001.
9. Hersh, W.R., Hickam, D.H. A Comparison of Retrieval Effectiveness for 3 Methods of Indexing Medical Literature. American Journal of the Medical Sciences. 303 (5): 292-300 May 1992.
10. Su, K.C, Ries, J.E, Peterson, G.M, Sievert, M.C, Patrick, T.B, Moxley, D.E, Ries, L.D. Comparing Frequency of Word Occurrences in Abstracts and Texts Using Two Stop Word Lists. Journal of the American Medical Informatics Association. 682-686 Suppl. S 2001.
11. Johnson, E.D., Sievert, M.C. and McKinin, E.J. Retrieving Research Studies: A Comparison of Bibliographic and Full-Text Versions of the *New England Journal of Medicine.* Proceedings of the Nineteenth Annual Symposium on Computer Applications in Medical Care, 1995, 846-850.
12. Srinivasan, P. MeSHmap: A Text Mining Tool for MEDLINE. Journal of the American Medical Informatics Association. 642-646 Suppl. S 2001.
13. Healey, P., Rothman, H., and Hoch, P. An Experiment in Science Mapping for Research Planning. Research Policy. Vol. 15. 1986.
14. Funk, M.E., and Reid, C.A. Indexing Consistency in MEDLINE. Bull Med Libr Assoc. 71 (2). 1983. 176–183.
15. Lancaster, F.W. *Vocabulary Control for Information Retrieval.* Information Resource Press, 1972.
16. Lindberg, D.A., Humphreys, B.L. Computer Systems that Understand Medical Meaning. In: Scherrer, J.R, Cote, R.A, Mandil, S.H, editors,

Computerized Natural Medical Language Processing for Knowledge Representation. Proceedings of the IFIP-IMIA WG6 International Working Conference; 1988 Sep 12-15; Geneva, Switzerland. Amsterdam: North-Holland; 1989. p. 5-17.
17. Humphreys, B.L, Schuyler, P.L. The Unified Medical Language System: Moving Beyond the Vocabulary of Bibliographic Retrieval. In: Broering NC. editor. High-Performance Medical Libraries: Advances in Information Management for the Virtual Era. Westport (CT): Meckler; 1993. p. 31-44.
18. Hersh, W.R, Hickam, D.H, Haynes, R.B, McKibbon, K.A. A Performance and Failure Analysis of Saphire with a Medline Test Collection. Journal of the American Medical Informatics Association. 1 (1): 51-60 Jan-Feb 1994.
19. Mijnhout, G.S, Hooft, L, van Tulder M.W, Deville, W.L.J.M, Teule, G.J.J, Hoekstra, O.S. How to Perform a Comprehensive Search for FDG-PET Literature. European Journal of Nuclear Medicine. 27 (1): 91-97 Jan 2000.
20. Kostoff, R. N., Tshiteya, R., Pfeil, K. M., and Humenik, J. A. Electrochemical Power Source Roadmaps using Bibliometrics and Database Tomography. Journal of Power Sources. 110:1. 163-176. 2002.
21. Kaiser, H.F. The Application Of Electronic Computers To Factor Analysis. Educational and Psychological Measurement. 20: 141-151. 1960.
22. Jackson, J. E. A Users Guide to Principal Components. Wiley, New York. 569. 1991.
23. Cattell, R.B. The Scree Test for the Number of Factors. Multivariate Behavioral Research. 1. 245 –276. 1966.
24. Kostoff, R. N. The Practice and Malpractice of Stemming. JASIST. 54:10. June 2003.

.

APPENDIX 1 – SUMMARY FACTOR DESCRIPTIONS

ANTIBODIES AND AUTOIMMUNE DISEASES

Different types of autoantibodies, especially anti-nuclear, anti-centromere, and extractable nuclear, and their relation to auto-immune diseases such as MCTD and SLE.

RAYNAUD'S SYNDROME-RELATED AUTOIMMUNE DISEASES

Auto-immune diseases associated with Raynaud’s Phenomenon, such as CREST syndrome.

SYSTEMIC LUPUS ERYTHEMATOSUS CLASSIFICATION

Compares specificity of the American Rheumatism Association criteria for the classification of systemic lupus erythematosus.

ADRENAL CORTEX HORMONES FOR LUPUS

Role of glucocorticoids in treating lupus erythematosus.

SCLEROTIC AUTOIMMUNE DISEASES

Scerloderma-spectrum autoimmune diseases, especially the CREST syndrome, and especially in females.

ENDOTHELIAL CELL ACTIVITY IN PSS

Endothelial cell activity in patients with progressive systemic sclerosis.

MIXED CONNECTIVE TISSUE DISEASE

Fatal conditions associated with mixed connective tissue disease.

CIRCULATING IMMUNE COMPLEXES

Serum levels of circulating immune complexes (including cryoglobulins) and immunoglobulins, especially IgG and IgM, with some emphasis on their relation to biliary cirrhosis.

DEFICIENCIES OF COMPLEMENT COMPONENTS

Hereditary deficiencies of complement components

FASCIA-FOCUSED INFLAMMATION

Inflammation, especially of the fascia (eosinophilic fasciitis), and the steroids used to control the inflammation.

CORONARY CIRCULATION AND HYPERTENSION

Coronary circulation and blood pressure problems; action of adrenergic antagonists, such as propranolol, on beta-2 adrenergic receptors.

SMOOTH MUSCLE RESPONSE TO DRUGS

In-vivo and in-vitro animal responses of smooth muscle to pharmacological agents, especially norepinephrine.

PLASMA LIPID FRACTION CONTROL

Prevention of myocardial infarction, angina pectoris, and cardiovascular arrhythmia through control of plasma lipid fractions.

CARDIAC BLOOD SUPPLY OBSTRUCTIONS

Cardiac and cerebral blood supply obstructions in young women, including migraine.

DOUBLE-BLIND VASODILATOR TRIALS

Double-blind trials for vasodilators such as nifedipine, or of imidazole derivatives as thromboxane-A synthase inhibitors for reduced platelet aggregation.

CALCIUM CHANNEL BLOCKERS

Calcium channel blockers and vasodilators, including their effect on smooth muscle contraction and treatment of non-cardiac disorders such as asthma.

BIOFEEDBACK TRAINING FOR IMPROVED CIRCULATION

Use of biofeedback and autogenic training techniques to induce relaxation, reduce stress headaches, and raise temperatures through improved circulation.

REDUCED PLATELET AGGREGATION VASODILATORS

Administration of vasodilators to improve circulation. Focuses on double-blind clinical trials of imidazole derivatives as thromboxane-A synthase inhibitors for reduced platelet aggregation. Also, therapeutic administration of drugs, mainly vasodilators such as prostaglandin-E, to increase regional blood supply.

PERIPHERAL CIRCULATORY SYSTEM VASODILATION

Vasodilation of the peripheral circulatory system after immersion, and the role of calcium in this process.

FINGER BLOOD FLOW MEASUREMENTS

Blood flow, and associated finger systolic blood pressure and temperature measurements.

NAIL-FOLD CAPILLARY MICROSCOPY

Diagnostic uses of electron microscopy and nail-fold capillary microscopy.

TRANSCUTANEOUS NERVE STIMULATION/ VASODILATION

Mediators of skin vasodilation induced by transcutaneous nerve stimulation.

SURGICAL/ NERVE BLOCK CIRCULATION OBSTRUCTION SOLUTIONS

Surgical and nerve block solutions to remove motor system constrictions on circulation, including surgical treatments for subclavian artery lesions and embolism at the thoracic outlet and other arteries.

CARPAL TUNNEL SYNDROME

Carpal tunnel syndrome.

VIBRATION-BASED CIRCULATION PROBLEMS

Impact of vibratory tools (such as chain saws) on circulation, and adverse effects of occupational vibration on the peripheral nervous system.

VINYL CHLORIDE EFFECTS ON BLOOD FLOW AND PROPERTIES

Vinyl chloride effects on blood flow and properties, especially chemically-induced diseases, such as osteolysis, resulting from the industrial use of vinyl chloride compounds.

TREATMENTS FOR CHEMICALLY-INDUCED NEOPLASMS

Chemotherapeutic treatments for neoplasms, especially testicular neoplasms.

Section 4. Systematic Acceleration of Radical Discovery and Innovation in Science and Technology

(based on Kostoff, R. N. "Systematic Acceleration of Radical Discovery and Innovation in Science and Technology". DTIC Technical Report Number ADA430720 (http://www.dtic.mil/). Defense Technical Information Center. Fort Belvoir, VA. 2005., and Kostoff, R.N. "Systematic Acceleration of Radical Discovery and Innovation in Science and Technology". Technological Forecasting and Social Change. 73 (8): 923-936. 2006.)

OVERVIEW

A systematic two-component approach (front-end component, back-end component) to bridging unconnected disciplines and accelerating potentially radical discovery and innovation (based wholly or partially on text mining procedures) is presented. The front-end component has similar objectives to those in the classical literature-based discovery (LBD) approach, although it is different mechanistically and operationally. The front-end component will systematically identify technical disciplines (and their associated leading experts) that are directly or indirectly-related to solving technical problems of high interest. The back-end component is actually a family of back-end techniques, only one of which shares the strictly literature-based analysis of the classical LBD approach. The non-LBD back-end techniques (literature-assisted discovery) make use of the human experts associated with the disparate literatures (disciplines) uncovered in the front-end to generate radical discovery and innovation.

Specifically, in the ***literature-assisted discovery*** operational mode, these disparate discipline experts could be used as:

1. Recipients of solicitation announcements (BAA, SBIR, MURI, journal Special Issue calls for papers, etc),
2. Participants in Workshops, Advisory Panels, Review Panels, Roadmaps, and War Games,
3. Points of Contact for Field Science Advisors, Foreign Field Offices, Program Officer site visits, and potential transitions

DEFINITIONS

Discovery is ascertaining something previously unknown or unrecognized. Innovation reflects the metamorphosis from present practice to some new, hopefully "better" practice. It can be based on existing non-implemented knowledge, discovery of previously unknown information, discovery and synthesis of publicly available knowledge whose independent segments have never been combined, and/ or invention. In turn, the invention could derive from logical exploitation of a knowledge base, and/ or from spontaneous creativity (e.g., Edisonian discoveries from trial and error). [1].

More generally, radical discovery and radical innovation depend on the source of the inspiration and/ or the magnitude of the impact. The more disparate the source of ideas from the target problem discipline, the more radical the potential discovery or innovation. The greater the magnitude of change/ impact resulting from the discovery or innovation, the more radical the potential discovery or innovation. The emphasis of the present section is on the breadth of the source.

BACKGROUND

Literature-based discovery (LBD) is a systematic two-component approach to bridging unconnected disciplines (front-end component, back-end component) based on text mining procedures. LBD allows potentially ***radical*** discovery and innovation to be hypothesized. Classically, the LBD front-end component has been used to identify the pool of potential discovery and innovation candidates, and the LBD back-end component has been used to hypothesize the potential discovery and innovation based on literature analysis alone.

The pioneering LBD study was reported in Swanson's paper hypothesizing treatments for Raynaud's Disease [2]. Subsequent LBD studies were performed by Swanson/ Swanson & Smalheiser on other medical problems [e.g., 3-6]. They also developed more formalized analytical techniques for hypothesizing radical discovery [e.g., 7-8]. Other researchers have used variants of Swanson's approach for hypothesizing radical discovery [e.g., 9-12]. Given:

- the length of time since Swanson's pioneering paper (two decades),
- the massive number of medical and technical problems in need of radical discovery, and

- the relatively few articles published in the literature using existing LBD approaches to generate radical discovery (especially articles not published by the Swanson/ Smalheiser team and not replicating the initial Raynaud's results),

it is clear that improvements in the fundamental approach and its dissemination and acceptability are required.

Additionally, LBD consists of literature analysis only for the entire process. Yet, the front-end of the process generates the pool of discovery candidates, including all types of documents and their authors, especially from disparate disciplines that historically serve as a powerful source of radical discovery. There appear to be no studies reported in the literature that have made explicit use of the discipline experts (identified in the front-end of the LBD process) for generating radical discovery and innovation.

INTRODUCTION

Discovery and innovation are the cornerstones of frontier research. One of the methods for generating radical discovery and innovation in a target discipline is to use principles and insights from disciplines very disparate to the target discipline, to solve problems in the target discipline. Unfortunately, identifying these linkages between the disparate and target disciplines, and making the subsequent extrapolations has tended to be a very serendipitous process. Until now, there has been no fully systematic approach to bridging these unconnected target and disparate disciplines. The present section describes a systematic approach, or more specifically, variants of a systematic approach (based wholly or partially on text mining procedures) for making these connections. One of the virtues of these specific approaches is that most of them can easily be integrated into the operational processes of science and technology (S&T) sponsoring organizations, or research performing organizations. However, some of these approaches do have characteristics of 'disruptive technologies', due to the additional effort required to properly integrate large numbers of concepts representing many disparate disciplines.

There are many examples where acceleration of discovery and/ or innovation require insights and knowledge from 'external' (i.e., indirectly-related or disparate) technical disciplines, sometimes very disparate disciplines. One could envision a solution to 'mine detection' that exploits

the remote detection of markers of nitrogen homeostasis in the presence of clinical disorders. In this case, the 'internal' (i.e., core, or directly related) technical discipline is that ordinarily associated with 'mine detection', while the 'external' technical disciplines would be those associated with specific aspects of remote (or possibly in-situ) detection not normally associated with 'mine detection'. The real challenge is to have a systematic process that identifies these 'external' disciplines starting from the 'internal' disciplines (thereby retaining some indirect thread of connectivity between the 'internal' and 'external' disciplines), and then extrapolates the insights and knowledge from these 'external' disciplines to solve problems in the 'internal' discipline or technology of interest.

The challenge has become more critical due to increasing specialization and effective isolation of technical/ medical researchers and developers [13]. As research funding and numbers of researchers have increased substantially over the past few decades, the technical literature has increased substantially as a result. Researchers/ developers struggle to keep pace with their own disciplines, much less to develop awareness of other disciplines. Thus, we have the paradox that the *expansion of research* has led to the *balkanization of research*! The resulting balkanization serves as a barrier to cross-discipline knowledge transfers, and retards the progress of discovery and innovation [13].

There are two main text mining avenues for extrapolating knowledge and insights from one discipline/ technology to another: ***literature-based discovery*** and ***literature-assisted discovery***. The ***literature-based discovery*** approach uses technical experts to access and examine the literature from 'external' disciplines to help solve problems in the 'internal' discipline. The ***literature-assisted discovery*** approach uses technical experts from 'external' disciplines in a variety of interactive and/ or independent creative modes for the same purpose.

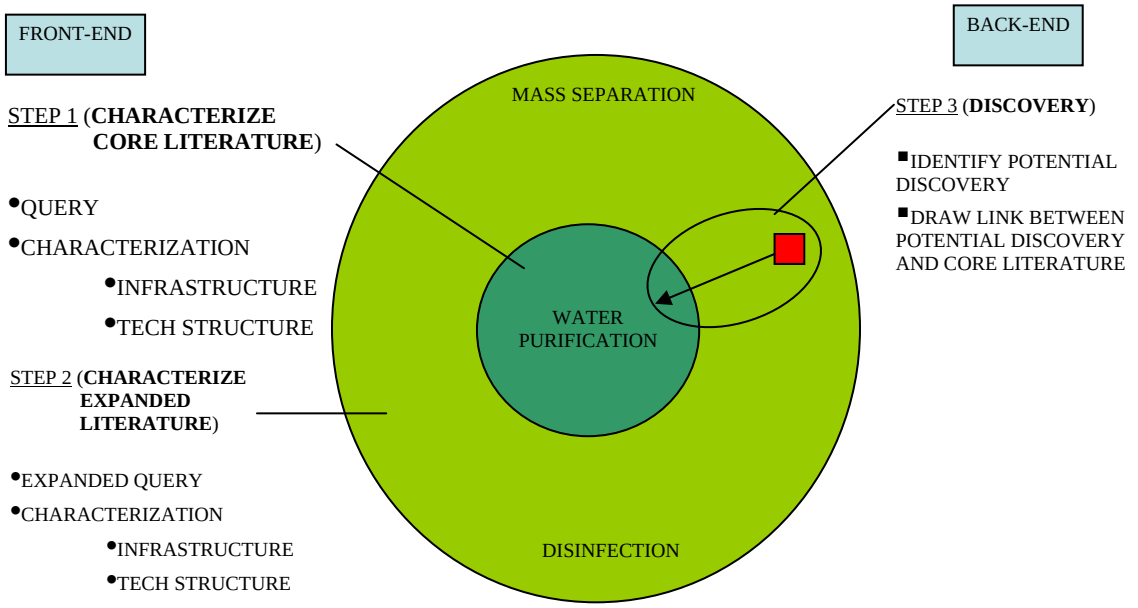
The main thesis of this section is that the scientific community has not made adequate use of these 'external' discipline sources of knowledge to accelerate potentially radical discovery and innovation. Further, very substantive quality enhancements to funding agency S&T programs, individual research projects, journal Special Issues, and multi-disciplinary teams and organizations are possible at relatively small marginal costs, if we can <u>systematically</u> improve access to the limitless sources of 'external' discipline/ technology information.

RADICAL DISCOVERY AND INNOVATION CONCEPT

Text mining is the extraction of useful information from large volumes of text [14-15]. The author's text mining effort over the past decade has been developing methods to systematically access external sources of information that could contribute to problem solving for specific technical disciplines, technologies, systems, operations, or technical problems in general [e.g., 16-25]. These methods have been integrated to form the following systematic approach for accelerating radical discovery and innovation.

Figure 1 contains a schematic of our text mining approach to discovery. The inner circle represents the core literature of the problem to be solved. In the example for Figure 1, the problem to be solved is identifying 'improved' alternatives to existing water purification technologies, where 'improved' could encompass any combination of lower cost, lower energy use, lower maintenance, higher reliability, lighter weight, and improved modularity for field assembly. Thus, the core literature is the existing (more or less commonly accepted) water purification literature. The annular region between the inner and outer circles represents literatures related more indirectly to the core literature.

Figure 1



The discovery process presented in the present section is divided into two components, a front-end and a back-end.

FRONT-END

Step 1

The front-end component (summarized to the left of the figure) contains two major steps: characterization of the core literature (Step 1), and characterization of the expanded literature, including identification of technical experts associated with this literature (Step 2). In Step 1, a query to retrieve the core literature is developed iteratively [e.g., 26-27]. Once the core literature has been retrieved with this query, it is subject to text mining [14-25]. Bibliometrics provides the technical infrastructure (key authors/ institutions/ countries/ journals, etc) of the core literature [e.g., 16-25], and computational linguistics provides the technical structure (technical thrusts, hierarchical taxonomies) of the core literature [e.g., 16-25]. Step 1 reflects the scope of many of our mono-technology text mining studies to date [e.g., 16-25].

The criticality of Step 1 cannot be overemphasized. The core literature represents the starting point for the expansion processes. The derived expanded literature determines the pool of discovery candidates. Any gaps in the core literature will be reflected as gaps in the pool of potential discovery candidates. Therefore, it is imperative that the core literature be as complete and comprehensive as possible for the discovery application.

References [26-27] above summarize the author's approach to query generation for core literature retrieval. Extensive exploitation of co-occurrence phenoma across many attributes is made. For discovery purposes in particular, techniques that specifically exploit the underlying semantic structure of the core literature should also be used, in addition to strictly co-occurrence techniques. The author has made extensive use of factor analysis in understanding the semantic/ conceptual structure of retrieved literatures, and has made less formal use of factor analysis for query refinement of the core literature. The factor matrix filtering technique [28] was developed to exploit the underlying semantic structure of a retrieved literature for the purpose of identifying high technical content phrases based on the strength of their contribution to semantic concepts. This is another approach for selecting new query terms.

More formal techniques that exploit the semantic structure, such as latent semantic indexing [10, 29], should also be examined for core literature definition. Whether these semantic structure exploitation techniques offer more than properly conducted attribute co-occurrence techniques [e.g., 27] remains to be demonstrated in practice. A multitude of information retrieval techniques have been examined for more than a decade at the TREC conferences [http://trec.nist.gov/pubs.html], and the interested reader is advised to examine the proceedings of these conferences.

Step 2

a. Overview

In Step 2, the query developed in Step 1 is generalized and expanded, again iteratively (see the next section for some practical techniques to generate the expanded query). This expanded query will retrieve records from literatures more indirectly related to the core literature. Insights and principles from these disparate literatures/ technical disciplines can be extrapolated to solve problems of the core literature. Thus, in the example on Figure 1, the core water purification literature query is expanded to cover/ retrieve all of mass separation and disinfection documents. Insights from very disparate mass separation and disinfection approaches can then be extrapolated to solve problems in water purification.

b. Details of Query Expansion

This section provides more detail on query expansion for discovery and innovation. The objective is to generalize the query terms while maintaining a delicate balance: the generalized terms bear some relation to the initial core literature retrieval terms while the relation is sufficiently indirect for the two literatures to be considered disjoint. Also, the expanded query terms are not overly general such that an unwieldy amount of data is retrieved, or the records retrieved are so distant from those of the core literature that impacts will be minimal [30].

The approach consists of examining each core literature query term, and testing different levels of generalizing the term to insure that the above objectives are met. If the records retrieved from the previous iteration are clustered, then terms for the expanded query should be selected such that each main theme from the clustering is adequately represented in the query [31]. Each term being considered for the expanded query should be tested in the source database (e.g., Science Citation Index [SCI]) for retrieval

efficiency of relevant records. In some cases, very general forms of the term should be inserted in the source database, and the retrieved records analyzed for more specific variants of the overly general form of the term.

At this point, an example may be illuminating. Consider the topics represented in Figure 1. The first precursor discovery objective is to expand the core water purification literature to include a more general indirectly-related literature containing potential discovery and innovation candidates. Clustering of the core literature may show very narrow and specific aspects of the assumed two main themes: distillation, membrane filtering ***(mass separation)***, ozonation, chlorination ***(disinfection)***. Phrases selected for the query should be drawn from each of the more generic versions of the main themes, shown bolded in the previous sentence.

For example, suppose WATER PURIFICATION is one of the query terms for retrieving the core literature. How could it be generalized for expansion, according to the principles set forth above? One approach is incremental generalization. WATER is a sub-set of LIQUID. Therefore, WATER PURIFICATION could be generalized incrementally to LIQUID PURIFICATION. Use of this query term in the source literature would retrieve documents on purification of liquids in addition to water, and any novel concepts used to purify liquids other than water could be extrapolated to help solve the water purification problem.

Before adding this more general term to the query, LIQUID PURIFICATION should be inserted into the SCI search engine, and the retrieval sampled to insure that a high fraction of records relevant to mass separation are being retrieved. In turn, LIQUID is a sub-set of FLUID. Therefore, LIQUID PURIFICATION could be generalized incrementally to FLUID PURIFICATION, and now concepts from the additional gas purification documents could be extrapolated to water purification improvements. The same following steps above would be repeated. The next generalization might be to MASS PURIFICATION, and so on.

How broadly should a query term be generalized? The more directly related the expanded literature is to the core literature, the more obvious will be the connections, but the lower will be the probability for radical discovery and innovation. The more indirectly related the expanded literature is to the core literature, the less obvious will be the connections, but the higher will be the probability for radical discovery and innovation. Thus, if radical discovery

and innovation is the goal, the broadest expansion consistent with available resources and reasonable numbers of links in the relational chain should be utilized.

For the WATER PURIFICATION query term being discussed, a second approach to generalization for expansion is to filter the core literature records for all phrases containing the word PURIFICATION. Then, each reasonable query expansion term candidate based on PURIFICATION can be checked following the steps above.

A third approach to generalization is to select one of the phrase words that is too generic to use as a stand-alone query term (e.g., WATER or PURIFICATION) for further examination. Since purification is the technology of interest, the word PURIFICATION would be inserted into the SCI search engine, and thousands of records retrieved. Text analyses would be performed on these retrieved records, and all phrases containing the word PURIFICATION would be extracted, and examined. Each PURIFICATION phrase variant proposed for the query would follow the same checks described above. While time consuming, this is the author's preferred approach for examining foundational terms for query expansion. In a water purification study, for example, this approach could be used for the very generic terms of foundational importance to the core separation processes (e.g., REMOVAL, SEPARATION, PURIFICATION, EXTRACTION, etc). This approach can provide quite comprehensive query terms. Because of the time involved for this latter expansion approach, only the most important generic roots should be examined. All other terms could be examined for expansion using the first or second methods described. As suggested in the core literature development section, once the expanded literature has been generated using the approaches of the preceding paragraph, it could be broadened further using techniques that exploit the underlying semantic structure [e.g., 10, 29].

How large should resultant queries be? For queries whose objective is retrieval of a statistically representative sample of documents from the source literature to define the core literature, our marginal utility approach can be used to determine cutoff [22]. Basically, more query terms provide increased refinement of a fixed scope literature (e.g., the existing water purification literature). In the reference example [22] (a text mining analysis of the NonLinear Dynamics literature), a 156 term query was reduced to 100 terms, with essentially no loss in fidelity of retrieval of the NonLinear

Dynamics literature. Other queries used in the past for this objective ranged from a handful of terms to hundreds of terms, depending strongly on the technology being examined.

For queries whose objective is retrieval of potential discovery items, most comprehensive retrieval coverage of an expanding scope literature, consistent with high retrieval precision (mainly relevant records), is required. The numbers of query terms will be higher than in the first case, and queries up to many hundreds of terms in length (depending on the specific technologies being studied) are possible, and in fact have been generated.

BACK-END

The back-end component contains the discovery step, which itself contains two sub-components. The first sub-component is identification of potential discovery and innovation candidates from the expanded literature, and the second sub-component is drawing the linkages between the potential discovery/ innovation candidates and the core literature. As will be shown in this section, there are many ways to identify potential discovery and innovation candidates, and to draw the subsequent linkages. These techniques differ mainly by the approach mechanics and the types of people used to identify the discovery and innovation candidates. The two main discovery and innovation approach types (Literature-Based, Literature-Assisted) are described now.

A. Literature-Based Discovery

We can use systematic techniques to identify potential discovery and innovation based strictly on the 'external' literature, an approach known as literature-based discovery (LBD) [2,8,9]. LBD is useful in the planning and concept identification phases of the S&T development cycle. The literature-based approach can be viewed as a very sophisticated type of literature survey, and represents a somewhat different way of doing business for most S&T sponsoring agencies, researchers, and technical journals. ***Done properly, LBD has the potential of generating at least an order of magnitude more true discovery than what has been reported in the LBD literature (as we are in the process of demonstrating).***

B. Literature-Assisted Discovery

We can identify technical experts associated with the 'external' indirectly-related disciplines, and then have them focus their expertise on solving problems of interest from the 'internal' disciplines. This literature-assisted people-based approach could easily be incorporated into most S&T sponsoring agencies' existing operational procedures. However, in some applications, proper handling of the infusion of large numbers of concepts and insights from disparate disciplines will acquire the characteristics of 'disruptive technologies'.

Thus, the differences between paths A and B above are in the 'back-end', in 1) how the linkages between the 'external' and 'internal' disciplines are made, and 2) who makes the linkages.

The ultimate goal should be incorporation of both approaches in parallel, to exploit the strengths of each approach while eliminating the weaknesses. This synergy would provide the *comprehensiveness and objectivity* of the completely literature-based approach coupled with the *interaction and feedback* of the literature-assisted people-based approach [1].

With respect to the literature-assisted approach, how would the 'external' discipline experts be incorporated into different components of the overall research enterprise's operations, for the purpose of enhancing and accelerating discovery and innovation? The following literature-assisted *options* are a sample of what is possible.

1) Solicitations – Science and Technology Sponsoring Organizations
Government agencies and private foundations generate numerous solicitations for proposals and/ or new ideas for solving problems. In the United States, these include Broad Agency Announcements (BAAs), Small Business Innovation Research (SBIR) solicitations, and other similar types. For the Federal agencies, these solicitations are usually advertised in some widely available forum (e.g., FedBizOpps) and, in parallel, some announcements of the solicitation are disseminated to technical experts deemed knowledgeable about the topic. ***These targeted announcements are extremely important, since they insure that the recipient will be aware of the solicitation***. The numbers of announcements are usually modest, because of limited prior access to potentially relevant communities.

OPTION 1. The *first option* is that the 'external' discipline experts identified through the text mining, as well as the comprehensive list of

'internal' discipline experts, constitute the bulk of the announcement distribution list. In this way, their expertise in a directly or indirectly-related discipline could be brought to bear on solving the sponsor's problem of interest, with their motivation amplified by the potential for funding, if successful. These 'internal' and 'external' discipline experts would also serve as the gateway to identifying additional technical experts (in their specific disciplines) not associated with the particular literatures accessed (e.g., through common professional societies, institution sub-divisions, attendance at conferences with the experts identified through the strictly literature approach), and thereby adding these additional technical experts to the problem-solving process. Use of the expanded notification list could result in an order of magnitude more proposals, and perhaps two orders of magnitude more potentially radical discovery and innovation proposals.

Potential consequences ('side-effects') resulting from this new approach to solicitations include 1) **substantial** increases in numbers of proposals (***as we have demonstrated successfully***), 2) need to expand diversity of reviewers' technical disciplines to insure interdisciplinary proposals receive balanced evaluation [13], and 3) need to facilitate/ stimulate discovery process by improved notification instructions. Consequences 1) and 2) could have modest 'disruptive technology' characteristics, especially if very large numbers of proposals from experts in many disparate disciplines are received.

2) Solicitations – Science and Technology Journals

Many technical specialty journals are structured on centuries-old research archival and dissemination models. Their scope is 'stove-piped' about quite narrow themes. This parochialism is further compounded by the increasing influence of Impact Factor as a publication metric target, restricting the types of articles published to increasingly narrower bands. The publication trend is toward more narrowly discipline-focused articles that will receive high citations. The trend is away from articles that encompass very diverse disciplines, may not receive high citations on average due to their inter-disciplinary nature [13], but could stimulate more radical discovery and innovation.

My recent citation studies of specific technical journals and of specific multi-journal technical disciplines show graphically the narrow dispersion of highly cited paper types, especially in the technical specialty journals. This trend needs to be reversed if the technical journals are to assume their

rightful positions as engines of innovation. The following paragraphs offer one approach for reversing this trend.

Most technical journals produce Special Issues periodically. These Special Issues tend to focus on a single topic, and usually have recognized experts on the specific topic present their perspectives. The papers tend to focus on comprehensiveness of coverage about the specific topic, rather then venturing into very disparate disciplines in a search for discovery.

There are two main avenues by which Special Issue papers are solicited. One is the Guest Editor (usually a recognized expert in the Special Issue topic) inviting other recognized experts known to him/ her. The second is the Guest Editor/ journal placing 'call for papers' ads in prior issues of the journal, or other closely-related topic-centric journals. In both cases, the result is the same: papers centered closely about the topic of interest.

However, the Special Issue concept could be expanded to emphasize radical discovery and innovation. As in the science and technology sponsoring organization example, notification of the projected Special Issue could be sent to the technical experts identified in the front end of the text mining discovery study. These experts would be encouraged to submit papers for the Special Issue that involved extrapolation of insights and principles from their own technical specialties to solving problems in the Special Issue topic. In this operational mode, the technical journals would serve proactively as the engines for radical discovery and innovation.

OPTION 2. The *second option* is that the 'external' discipline experts identified through the text mining, as well as the comprehensive list of 'internal' discipline experts, constitute the bulk of the journal Special Issue announcement distribution list. In this way, their expertise in a directly or indirectly-related discipline could be brought to bear on solving the Special Issue's particular problem of interest, with their motivation amplified by the potential for journal publication, if successful.

The potential consequences from this new mode of Special Issue operation mirror those of the first option: more papers than normal, greater topical diversity in papers, and need to facilitate discovery and innovation. It is strongly recommended that the Special Issue be expanded in size from the normal journal practice, to accommodate the expected increase in number of submittals. Sponsorship of such Special Issues by the science and

technology funding organizations seems appropriate. Modest honoraria could also be provided to authors as further motivation for participation.

3) Advisory Panels
Government agencies and private organizations convene numerous advisory panels or groups of independent advisors for the purpose of providing expert technical advice on problems of present and future interest. Unfortunately, many of these advisory groups are somewhat parochial, both in terms of technical scope and people, thereby limiting the breadth of their recommendations.

OPTION 3. The *third option* is that the 'external' (and 'internal') discipline experts identified through the text mining constitute a significant portion of the members of these groups.

4) Workshops
Government agencies and private organizations conduct numerous workshops for the purpose of generating new project ideas and directions. Many of these workshop participants are somewhat parochial, both in terms of technical scope and people. Additionally, venture capital organizations, and other components of Wall Street, have great needs for workshops that could provide insight on the potential of emerging technologies. It would be useful for these organizations to know whether the core technology of interest is amenable to improvement, and whether some of the indirectly-related technologies can offer potential solutions across many different core technologies.

OPTION 4. The *fourth option* is that the 'external' (and 'internal') discipline experts identified through the text mining constitute a significant portion of the participants at these workshops [1]. If the workshops are conducted in tandem with the solicitation processes above, then the results from the solicitations could be used to narrow the pool of candidates for these workshops. The workshop attendees could be drawn from the solicitation announcement recipient group that submitted proposals to the sponsoring organization solicitations, or the solicitation announcement recipient group that submitted papers to the technical journal solicitations. At a minimum, the members of these groups had sufficient insight to perceive how concepts from their areas of expertise could be extrapolated to solve problems of interest in the core technology.

5) Review Panels
Government agencies and private organizations conduct numerous review panels during the execution of their S&T programs. Some agencies use such panels to help evaluate proposals and provide recommendations. Many of these review/ evaluation panels are somewhat parochial, both in terms of technical scope and people. This limits discussion of the breadth of approaches that could be used to achieve the program objectives.

OPTION 5. The *fifth option* is that a portion of the reviewers be drawn from the 'external' (and 'internal') discipline experts identified through the text mining.

6) Roadmaps
Some government agency and private organization programs generate technology roadmaps (see [32] for a description of technology roadmaps) as part of their planning processes, and/ or as part of their review processes. Many of these roadmap development teams have limited perspectives, both in terms of technical scope and people. The breadth of these roadmaps is limited by the breadth of their developers.

OPTION 6. The *sixth option* is that a sub-set of the 'external' (and 'internal') discipline experts identified through the text mining constitute a significant portion of the roadmap development team. As in the workshop option, the solicitation step could be used to filter the candidate pool for prospective roadmap development team members.

7) Points of Contact
Many government agency components require points of contact (POCs) for obtaining information to solve problems. These include the Field Science Advisors for military organizations, Foreign Field Offices of government agencies, Program Officers for site visits, and Program Officers to identify potential transitions. In practice, many of these POCs accessed derive from limited personal knowledge, both in terms of technical scope and people. The breadth of the information obtained from the POCs is limited by their breadth of expertise.

OPTION 7. The *seventh option* is that the 'external' (and 'internal') discipline experts identified through the text mining constitute a very significant portion of the POCs accessed in practice. Again, as in the

workshop option, the solicitation step could be used to filter the candidate pool for prospective POCs.

8) Organization and Team Structuring
Technical teams and organizations can be structured to maximize the potential for radical discovery and innovation, based on the principles presented in the paragraphs above. Multi-disciplinary groups/ structures such as Integrated Product Teams (IPTs), Multi-Disciplinary Research Programs of the URI (MURIs), and Cooperative Research and Development Agreements (CRADAs) could be assembled based on the technical disciplines and technical experts identified at the front-end of the discovery process above. Organizations such as Centers of Excellence with a defined core competency, and large laboratories with multiple core competencies, could be structured to incorporate the technical disciplines surrounding their core identified at the front-end of the discovery process.

OPTION 8. The *eighth option* is that the 'external' (and 'internal') disciplines identified through the text mining constitute a significant portion of the teams and organizations. Where possible, the solicitation step could be used to filter the candidate pool for prospective team and organization members.

9. War Games
War games model and/ or simulate different pre-combat, combat, and post-combat scenarios. Each Title 10 war game, in addition to major and minor players, the assessors, and the game controlling authority, also has various supporting cells to provide expertise and insight in key areas. For example, the Science and Technology Cell's primary function at the *Global 98* game was to provide future projections of technology that could be available to the game players, and to suggest applications resulting from ongoing R&D. The effectiveness of the advice from such S&T cells depends on the breadth of technical understanding of the cells' members. Additionally, the value of the war games depends in part on the technology capabilities designed into the games and the post-games evaluation of the role that technology played in impacting tactics and strategy.

OPTION 9. The *ninth option* is that the 'external' (and 'internal') disciplines identified through the text mining constitute a portion of the supporting S&T cells' memberships, the membership of the games'

technology designers, and the membership of the post-games' technology evaluators.

SUMMARY AND CONCLUSIONS

I have proposed the identification and exploitation of diverse literatures, and their representative experts, to help solve problems of interest through potentially radical discovery and innovation. The approach is based on our demonstrated text mining techniques. Their essential element is development of comprehensive and precise queries for retrieving the expanded literature of potential discovery candidates, followed by exploitation of these retrieved literatures and their associated technical expert representatives.

I have identified a number of pathways by which these literatures and people could be integrated with present business practices, or could be integrated with slight modifications if desired. ***This group of 'external' literature accession techniques has the highest benefit/ cost ratio of any techniques I know for enhancing and accelerating radical discovery and innovation.***

Additionally, the "structural holes" research of Professor Burt [33], which identifies 'structural holes' as the weakly-linked or non-linked region between different technical disciplines, has shown that most innovations studied have been the result of drawing on insights from unconnected, sometimes very disparate, disciplines/ technologies. These findings support the thesis that is the basis for my proposal above. Translated into practice, the following important guidelines can be drawn.

1. If we are interested in meeting short-term deadlines efficiently with specific well-defined technology products, then homogeneous well-coordinated and long-standing groups are most useful.
2. If we are interested in radical discovery and innovation, including innovation in the advanced technology demonstration of system integration sense, then we need to incorporate people *we don't know* representing disciplines with which *we are not familiar*. It is difficult to get ***'out-of-the-box'*** thinking from people who have spent their careers ***'in-the-box'***!

The generic literature-based and literature-assisted discovery approaches provide a systematic and objective guide to selecting the most appropriate

disciplines and people for accelerating potentially radical discovery and innovation.

REFERENCES FOR SECTION 4.

1. Kostoff RN. Stimulating innovation. *International Handbook of Innovation*. Larisa V. Shavinina (ed.). Elsevier Social and Behavioral Sciences, Oxford, UK. 388-400. 2003.
2. Swanson DR. Fish oil, Raynauds syndrome, and undiscovered public knowledge. Perspectives in Biology and Medicine. 30 (1): 7-18 Fall 1986
3. Swanson DR. Mmigraine and magnesium - 11 neglected connections. Perspectives in Biology and Medicine 31 (4): 526-557. Summer 1988
4. Swanson DR. Somatomedin-C and arginine - implicit connections between mutually isolated literatures. Perspectives in Biology and Medicine. 33 (2): 157-186 Winter 1990.
5. Smalheiser NR, Swanson DR. Assessing a gap in the biomedical literature - magnesium-deficiency and neurologic disease. Neuroscience Research Communications 15 (1): 1-9. July-August 1994.
6. Swanson DR, Smalheiser NR, Bookstein A. Information discovery from complementary literatures: Categorizing viruses as potential weapons Journal of the American Society for Information Science and Technology. 52 (10): 797-812. August 2001.
7. Swanson DR, Smalheiser NR. An interactive system for finding complementary literatures: a stimulus to scientific discovery. *Artificial Intelligence*. 91(2). 1997.
8. Smalheiser NR, Swanson DR. Using ARROWSMITH: a computer-assisted approach to formulating and assessing scientific hypotheses Computer Methods and Programs in Biomedicine 57 (3): 149-153. Nov 1998.
9. Gordon MD, Lindsay RK. Toward discovery support systems: A replication, re-examination, and extension of Swanson's work on literature-based discovery of a connection between Raynaud's and fish oil . Journal of the American Society for Information Science 47 (2): 116-128. Feb 1996.
10. Gordon MD, Dumais S. Using latent semantic indexing for literature based discovery. Journal of the American Society for Information Science 49 (8): 674-685. June 1998.
11. Weeber M, Klein H, de Jong-van den Berg LTW, et al. Using concepts in literature-based discovery: Simulating Swanson's Raynaud-fish oil and migraine-magnesium discoveries. Journal of the American Society for Information Science and Technology. 52 (7): 548-557. May 2001.

12. Stegmann J, Grohmann G. Hypothesis generation guided by co-word clustering. Scientometrics 56 (1): 111-135. 2003.
13. Kostoff, RN. Overcoming specialization. *BioScience*. 52:10. 937-941. 2002.
14. Kostoff, RN. Text mining for global technology watch. In Encyclopedia of Library and Information Science, Second Edition. Drake, M., Ed. Marcel Dekker, Inc. New York, NY. 2003. Vol. 4. 2789-2799.
15. Hearst, M. Untangling text data mining. In the Proceedings of ACL'99: the 37th Annual Meeting of the Association for Computational Linguistics, University of Maryland, June 20-26, 1999.
16. Kostoff RN, Eberhart HJ, and Toothman DR. Database tomography for technical intelligence: a roadmap of the near-earth space science and technology literature. Information Processing and Management. 34:1. 69-85. 1998.
17. Kostoff RN, Eberhart HJ, and Toothman DR. Hypersonic and supersonic flow roadmaps using bibliometrics and database tomography. Journal of the American Society for Information Science. 50:5. 427-447. 15 April 1999.
18. Kostoff RN, Braun T., Schubert A, Toothman DR., and Humenik JA. Fullerene roadmaps using bibliometrics and database tomography. Journal of Chemical Information and Computer Science. 40:1. 19-39. Jan-Feb 2000.
19. Kostoff RN, Green KA, Toothman DR, and Humenik J. Database tomography applied to an aircraft science and technology investment strategy. Journal of Aircraft, 37:4. 727-730. July-August 2000.
20. Kostoff RN, and DeMarco RA. Science and technology text mining. Analytical Chemistry. 73:13. 370-378A. 1 July 2001.
21. Kostoff RN, Tshiteya R, Pfeil KM, and Humenik JA. Electrochemical power source roadmaps using bibliometrics and database tomography. Journal of Power Sources. 110:1. 163-176. 2002.
22. Kostoff RN, Shlesinger M, and Tshiteya R. Nonlinear dynamics roadmaps using bibliometrics and database tomography. International Journal of Bifurcation and Chaos. 14:1. 61-92. January 2004.
23. Kostoff RN, Shlesinger M, and Malpohl G. Fractals roadmaps using bibliometrics and database tomography. Fractals. 12:1. 1-16. March 2004.
24. Kostoff RN, Karpouzian G, and Malpohl G. Text mining the global abrupt wing stall literature. Journal of Aircraft. 42:3. 661-664. 2005.

25. Kostoff RN, Tshiteya R, Pfeil KM, Humenik JA., and Karypis G. Power source roadmaps using database tomography and bibliometrics. Energy. 30:5. 709-730. 2005.
26. Kostoff RN, Eberhart HJ, and Toothman DR. Database Tomography for information retrieval. Journal of Information Science. 23:4. 1997.
27. Kostoff RN. Systematic acceleration of radical discovery and innovation in science and technology. DTIC Technical Report Number ADA430720 (http://www.dtic.mil/). Defense Technical Information Center. Fort Belvoir, VA. 2005.
28. Kostoff RN, and Block JA. Factor matrix text filtering and clustering. JASIST. 56:9. 946-968. July. 2005.
29. Deerwester S, Dumais ST, Furnas GW, Landauer TK, Harshman R. Indexing by latent semantic analysis. Journal of the American Society for Information Science 41 (6): 391-407. Sept 1990.
30. Kostoff RN. Research impact quantification. R&D Management. 24:3. July 1994.
31. Kostoff RN. Method for data and text mining and literature-based discovery. United States Patent Number 6,886,010. 26 April 2005.
32. Kostoff RN, and Schaller RR. Science and technology roadmaps. IEEE Transactions on Engineering Management. 48:2. 132-143. May 2001.
33. Burt RS. Structural holes: The social structure of competition. Harvard University Press. Cambridge, MA. 1992.

Section 5. Overcoming Specialization

(based on Kostoff, R. N. "Overcoming Specialization." BioScience. 52:10. 937-941. 2002.)

OVERVIEW

This section describes the use of modern information technology to identify the balance of research disciplines required to comprehensively address any research problem, including when multi-disciplinary or inter-disciplinary approaches should be used.

INTRODUCTION

Many of the most challenging problems in bio-science require advances in a multitude of technical and non-technical disciplines in order for progress to be made. Bio-diversity, bio-complexity, bio-technology, species and eco-system conservation, and bio-terrorism require expertise from myriad technical, legal, political, financial, and cultural disciplines.

For example, environmental bio-complexity is founded on the idea that research on the individual components of environmental systems provides only limited information about the behavior of the systems themselves. Careful attention to the interplay among components is critical to obtaining the level of predictive information on which management and regulatory decisions must be made.

To understand the complex inter-dependencies among living organisms and the environments that affect, sustain, and are modified by them, efforts that transcend multiple disciplines are required that: span temporal and spatial scales, consider multiple levels of biological organization, cross conceptual boundaries, use contemporary technologies, and link research to environmental decision-making. Advancing understanding of the nature and role of biological complexity demands increased attention and new collaborations of scientists from a broad spectrum of fields -- biology, physics, chemistry, geology, hydrology, statistics, engineering, computation, and social sciences. In addition, advances in large-scale applications of remote sensors to monitor environmental bio-systems require access to the latest science and engineering literature in remote sensing, non-destructive evaluation, signal and image processing, pattern recognition, multi-source

data fusion, fluid dynamics, acoustics, robotics, materials, electronics, and many other disciplines.

Thus, in complex bio-science problems, addressing only one or a few of the component disciplines will result in fragmented or perhaps misleading results due to neglect of discipline inter-dependencies. However, even if the many disciplinary facets of a complex bio-science problem are addressed, the method of integration of the multiple facets can impact the solution of the problem. Research that includes multiple disciplines but maintains their distinctiveness is multi-disciplinary (Collins 2002). Such research may not include joint planning, management, and review of the multiple disciplines. Research that integrates the multiple disciplines to effectively form a new unified discipline is inter-disciplinary. Even if all of the multiple component disciplines are addressed separately in a multi-disciplinary approach, the final research product will not have the same quality as a unified research product resulting from an inter-disciplinary study, especially if the different disciplines impact each other strongly.

Another strong motivation for examining multiple disciplines is increased evidence that there are common underlying themes across many research fields. For example, the same equations are used to model phenomena in some very diverse disciplines, such as the modeling of chaotic behavior. Appropriate inter-discipline research and information transfer can allow findings and insights from one discipline to be extrapolated and exploited by another, perhaps very disparate, discipline.

Paradoxically, in parallel with the increasing need for inter-disciplinary projects, researchers have become much more specialized by necessity. The massive global expansion of technical literatures and other science and technology products reduces the time available for researchers to remain current in their own specialty disciplines, much less to become familiar with progress in other disciplines. In addition to lack of time, they also have many other dis-incentives to participate in inter-disciplinary projects (see Box 1). If there are no external incentives offered for inter-disciplinary research, most researchers will take the path of least resistance, and restrict their research projects within their own, or very closely related, disciplines.

In recent years, research sponsoring agencies have decided there is merit to inter-disciplinary research, and have provided incentives for the proposal and establishment of such programs. In many cases, the result has been

programs that are inter-disciplinary on paper only. They are not managed or reviewed as a cohesive inter-disciplinary unit, but are managed and reviewed (in practice) as fragmented separate programs. In other cases, programs (and facilities) have been advertised as inter-disciplinary when in reality each 'discipline' is a minor variant of a single discipline (e.g., Physics/ Materials, where the materials group members are basically physicists who happen to be focusing on the physics of materials). The number of true inter-disciplinary projects and programs that incorporate distinctly different disciplines, but are selected, managed, reviewed, and transitioned as cohesive units, is a small percentage of all research conducted.

Further, it is difficult to objectively gauge the effectiveness of these inter-disciplinary efforts. The metrics used for these assessments, such as numbers of paper authors from different disciplines or mixes of discipline funding under program managers, are very incomplete. These quantitative metrics are amenable to manipulation, can be deceptive, and intrinsically do not describe the quality of the discipline mixing process. Most egregiously, they do not separate artificial inter-disciplinary projects, such as the Physics/ Materials example above, from coherent projects consisting of relatively disparate disciplines.

However, it is not necessary to conduct all research programs as inter-disciplinary. There are some tangible and intangible costs involved in conducting inter-disciplinary programs, due to the overhead required to integrate diverse technical cultures and traditions (see Box 1). A program should be conducted as inter-disciplinary only if a strong diverse mix of disciplines is required to fully address its research objectives. There is no intrinsic virtue to conducting projects or programs as inter-disciplinary, unless it can be demonstrated that they fundamentally require an inter-disciplinary approach for maximum advancement.

PROCESS CONCEPT

The fundamental thesis of this section is that the mix of disciplines used in the conduct of a science and technology program should correspond to the multiple discipline requirements of the program. I propose a systematic three-step process (based on the use of modern information technology) for determining the relationship of the disciplines required to conduct a science and technology program to the disciplines selected. The first step in the process is *identification of the multiple disciplines* that could have some

impact on the research problem. The second step is *determination of the cost-effectiveness* (importance versus costs) of employing all the disciplines that could potentially impact the problem. The third step is *provision of incentives/ mandates* to the performers for incorporating those required disciplines that will contribute to the problem's solution cost-effectively.

PROCESS MECHANICS

Background

The proposed three-step process is based on text mining, and uses some of the concepts presented in section 4. For ease of comprehension, some of the material from section 4 will be restated, where required.

Text mining is the extraction of useful information from large volumes of text (Hearst 1999, Kostoff and DeMarco 2001a, Kostoff 2002a). Typically, text mining uses computational linguistics (e.g., phrase occurrence and co-occurrence frequencies) and bibliometrics (e.g., author, journal, and institution occurrence and co-occurrence frequencies) coupled with expert human judgement, to extract useful information from unstructured (free text) and semi-structured text (e.g., author, journal, and address fields). Extraction of the technical phrases and their occurrence frequencies from text identifies the pervasive science and technology areas within the text. Extraction of the phrase co-occurrences within some domain (e.g., Abstract, paragraph) provides the relationships among the science and technology areas, and provides the foundation for identifying new relations among allied and disparate science and technology areas.

For the past decade, one of the components of text mining known as literature-based discovery (Swanson 1986, Swanson and Smalheiser 1997, Gordon and Lindsay 1996, Weeber et al 2001, Kostoff 1999, Kostoff 2002b) has been used to identify, retrieve, and integrate appropriate disparate literatures for the purpose of generating innovation (see Box 2 for a more detailed description of literature-based discovery). In literature-based discovery, identification and merging of concepts from very different technical disciplines is not an option; it is a requirement.

The literature-based discovery studies that have been performed confirm the parochialism of researchers in the specific disciplines studied. Consider Swanson's initial paper on literature-based discovery (Swanson 1986), in

which he hypothesized that Fish Oil/ Eicosapentaenoic Acid could alleviate some symptoms of Raynaud's Disease (later confirmed by laboratory and clinical tests). The Raynaud's Disease researchers were not aware (based on what could be deduced from the literature analysis) of the Fish Oil literature, and the Fish Oil researchers were not aware of the Raynaud's literature.

Further, a 2001 bio-terrorism-related literature-based discovery study (Swanson et al 2001) identified viruses that are not recognized today as bio-warfare agents, but have the characteristics to be modified into bio-warfare agents. Such viral agents pose a special threat, since their use would contain the element of surprise. For such agents, there would be no vaccines for prevention, no detection, and perhaps no therapies, and the potential destructive consequences would be far greater than those of the anthrax bacterium. These viruses had gone un-recognized as candidate bio-warfare agents by the technical specialty communities. The two main bio-warfare agent characteristics, virus pathogenicity and virus transmissibility, had been studied by two disjoint research communities that were not familiar with each other's literatures (based on what could be deduced from the literature analysis).

First and Second Steps

The first step in the process is to perform a literature-based discovery analysis of the research problem prior to initiation of a research project. The output would consist of identifying: 1) technical disciplines that could potentially contribute to advances in the research problem; 2) experts within these disciplines; and possibly (not necessarily) 3) potential problem solutions.

In the tandem second step, the proposers or principal investigators could then estimate the importance of each of the identified disciplines to the attainment of the research objectives, and use that as the basis for a strategy of constructing the research approach.

This second step would use the output from the literature-based discovery for convening workshops or groups of experts (Kostoff 2002b). In contrast to standard workshops (see characteristics below), these workshops would be guided, where facilitators would actively enhance the transfer of cross-discipline information. The combination of literature-based discovery

followed by guided workshops would eliminate the deficiencies of standard workshops:

1) Small community representation
2) Parochialism; not all relevant disciplines represented
3) Human dynamics; can overwhelm technical discussions
4) High degree of subjectivity

as well as the deficiencies of literature-based discovery:

1) Only a small fraction of R&D conducted gets published
2) Currency; there is a lag time in publication
3) Minimal human interaction for concept stimulation
4) A specific solution to the problem may not be identifiable from the literature alone

This combination would retain the strengths of each component to produce a systematic enhancement of the environment for stimulating innovation. In the workshop, the range of required disciplines would be clarified further, and disciplines added or subtracted to the proposed research approach as dictated by the additional costs and benefits to science and technology. In addition, if the literature-based discovery has generated discovery in the form of specific hypotheses to be tested, these could be discussed and sharpened further.

An initial experiment was performed of this hybrid approach (Kostoff 1999), on the topic of Autonomous Flying Systems. A broad-based literature survey was performed, focused more on retrieval than discovery, and experts were identified from many disciplines that had some common thread with Autonomous Flying Systems. Experts selected for the workshop were asked to identify emerging opportunities from their disciplines well before the actual workshop, and then cross-discipline relationships of these opportunities were amplified through facilitated pre-meeting internet communications. The workshop meeting time was then used efficiently to focus on the most promising cross-discipline relationships and transfers.

The results appeared very positive. However, it became clear that more development of the literature-based discovery process was required to insure that the most comprehensive identification of potentially relevant disciplines was made. This is important for identifying, at the workshop, solutions to

the problem of interest that might not have been identified from the literature alone.

Third Step

The first two steps are mechanistic technology steps. They will work technically, although improvements in each are desirable and possible. The third step is the most difficult, since it involves incentives and the accompanying human issues of motivation, tradition, culture, and inertia. If progress is to be made in pursuing intrinsically inter-disciplinary research appropriately, mandates requiring at least the first step of the hybrid process (literature-based discovery) are probably required initially. After the technical community becomes convinced of the benefits of incorporating text mining at the initiation of research projects, and becomes familiar with the process mechanics involved, then incentives can probably replace mandates for performing pre-project text mining.

There is precedent for these types of pre-project literature survey mandates. A number of Federal agencies require literature surveys before initiation of research projects. Since text mining (sans workshop) could be viewed as a sophisticated form of literature survey, introduction of a pre-project text mining requirement would in some sense be an extension of existing literature survey requirements.

SUMMARY AND CONCLUSIONS

A three-step process has been proposed for insuring selection of a comprehensive mix of research disciplines to address a research problem. The process is based on the text mining variant of literature-based discovery to identify and select the comprehensive discipline mix before research is started. When appropriate, workshops can be convened using the information developed in the literature-based discovery component.

In this scenario, the literature-based discovery approach would serve as one block in the foundation of all research performed, in helping to objectively determine the mix of disciplines required to attain the research objectives. It may also provide discovery based on the literature studies alone. Even if actual discovery does not result from the literature phase alone, the fundamental value of literature-based discovery in determining discipline

mixes for subsequent workshops and research program conduct remains undiminished.

To insure that most of the potentially important disciplines are identified by the literature-based discovery process, more process development is required, and more variants of literature-based discovery are required. The quality and credibility of the literature-based discovery output depends on:

1) Study objectives; metrics used
2) Source databases used (e.g., Medline, Science Citation Index, Pascal)
3) Information retrieval techniques used
4) Record fields analyzed (e.g., Keywords, Titles, Abstracts, Full Text)
5) Analysis techniques, especially co-occurrence and clustering techniques (Kostoff et al 2001b)
6) Most importantly, the people performing the analysis

Each variant of literature-based discovery will use one or more alternatives of these study components, and only very few literature-based discovery studies have been published so far. This expanded development of literature-based discovery has not yet been started, and the discipline is one that has *completely fallen through the cracks* relative to government and industry funding.

This deficiency is particularly egregious relative to the present global threat from bio-terrorism. To the author's knowledge, Swanson et al (2001) was the only published text mining study to have addressed bio-warfare agent prediction. *One small study, using one approach, represents the total reported global text mining effort to prevent surprise by potential biowarfare agents that could be identified with publically available knowledge!* In what other area of science and technology is only one approach, no matter how good, used to solve a problem? Multiple literature-based discovery approaches, and multiple studies, are required to insure that as many candidate bio-warfare agents as possible are identified.

A national effort is needed to develop parallel literature-based discovery approaches, to insure that optimal methods are used to identify and integrate findings from disparate disciplines. Further, experiments are required to identify how the literature-based discovery results should be integrated with workshops to exploit these multi-disciplinary findings and maximize the potential for innovation. Finally, the requirement for incorporating

literature-based discovery at the initiation of research projects, to insure that all relevant research reported and all potentially relevant disciplines are identified, should be mandated for all Federally-supported research. Such a process would identify research that required multiple disciplines for rapid advancement, as well as research that could produce acceptable results from mono-discipline analysis.

REFERENCES FOR SECTION 5

Bauer HH. 1990. Barriers against interdisciplinarity-implications for study of science, technology, and society (STS). Science, Technology, and Human Values. 15:1. 105-119.

Bruhn JG. 1995. Beyond discipline: creating a culture for interdisciplinary research. Integrative Physiological and Behavioral Science. 30:4. 331-341.

Butler D. 1998. Interdisciplinary research 'being stifled'. Nature. 396. 19 November. 202.

Collins JR. 2002. May you live in interesting times: using multidisciplinary and interdisciplinary programs to cope with change in the life sciences. BioScience. 52:1. 75-83.

Geissler E., ed. 1986. *Biological and Toxin Weapons Today*. SIPRI: Oxford University Press.

Gordon MD, Lindsay RK. 1996. Toward discovery support systems: a replication, re-examination, and extension of Swanson's work on literature-based discovery of a connection between Raynaud's and fish oil. Journal of the American Society for Information Science. 47: 2. 116-128.

Hearst MA. 1999. Untangling text data mining. Proceedings of ACL 99, the 37th Annual Meeting of the Association for Computational Linguistics. University of Maryland. June 20-26. 1-9.

Kostoff RN. 1997. Peer review: the appropriate GPRA metric for research. Science. 277. 1 August. 651-652.

Kostoff RN. 1999. Science and technology innovation. Technovation. 19:10. 593-604.

Kostoff, RN, DeMarco RA. 2001a. Science and technology text mining. Analytical Chemistry. 73:13. 370-378A.

Kostoff RN, Del Rio JA, García, EO, Ramírez AM, Humenik JA. 2001b. Citation mining: integrating text mining and bibliometrics for research user profiling. Journal of the American Society for Information Science. 52:13. 1148-1156.

Kostoff, RN, Hartley J. 2001c. Structured abstracts for technical journals. Science. 5519:1067a. 292.

Kostoff RN. 2002a. Text mining for global technology watch. Encyclopedia of Library and Information Science. In Press.

Kostoff, RN. 2002b. Stimulating innovation. International Handbook of Innovation. In Press.

Metzger N, Zare RN. 1999. Interdisciplinary research: from belief to reality. Science. 283. 642-643.

Naiman RJ. 1999. A perspective on interdisciplinary science. Ecosystems. 2. 292-295.

Swanson DR. 1986. Fish oil, Raynauds syndrome, and undiscovered public knowledge. Perspectives in Biology and Medicine. 30: 1. 7-18.

Swanson DR, Smalheiser NR. 1997. An interactive system for finding complementary literatures: a stimulus to scientific discovery. Artificial Intelligence. 91:2. 183-203.

Swanson DR, Smalheiser NR, Bookstein A. 2001. Information discovery from complementary literatures: categorizing viruses as potential weapons. Journal of the American Society for Information Science and Technology. 52: 10. 797-812.

Weeber M, Klein H, de Jong-van den Berg LTW, Vos R. 2001. Using concepts in literature-based discovery: simulating Swanson's Raynaud-fish oil and migraine-magnesium discoveries. Journal of the American Society for Information Science and Technology. 52:7. 548-557.

BOX 1 – MULTI-DISCIPLINARY AND INTER-DISCIPLINARY RESEARCH BARRIERS

Some of the specific barriers to multi-disciplinary and inter-disciplinary research include Culture, Time, Evaluation, Publication, Employment, Funding, Promotion, and Recognition.

Culture

Different technical disciplines represent different cultures and traditions. Each culture has its own vocabulary, its own perspective on what constitutes evidence, its own standards of proof, its own definitions of truth, and its own traditions on how research is defined and performed. Merging of cultures and traditions for inter-disciplinary research requires communication, coordination, and consensus among cultures, and compromise from all parties. Additional time is required to structure inter-disciplinary proposals, and to plan the conduct of research projects (Bauer 1990, Naiman 1999).

Time

Inter-disciplinary research requires that each participant learn some aspects of the other participants' disciplines, including the cultures and traditions noted above. Time is required to learn these other technologies, cultures, traditions, and to effect the coordination and consensus processes. This time expenditure detracts from time spent on the mastery of a single discipline (Naiman 1999).

Evaluation

Peer review is the main and preferred type of research evaluation (Kostoff 1997). Traditionally, peer review has consisted mainly of judgements from mono-discipline reviewers, often in the same research area as the reviewee (Bruhn 1995, Metzger and Zare 1999, Butler 1998). Reviewers tend to give higher marks to in-depth advances made in a single discipline rather than less intense advances made across a wider range of disciplines.

Publication

Most ranked journals tend to have a strong mono-disciplinary mission, and many will even discourage submittal of broader-based inter-disciplinary manuscripts (Bruhn 1995, Butler 1998, Naiman 1999). The manuscript review process tends to have similar structure and reviewer parochialism problems for inter-disciplinary research as noted above under Evaluation.

The document Abstract, the main vehicle for communicating research content across disciplines in the large databases such as Medline and Science Citation Index, is in many cases incomprehensible to all but the research area experts (Kostoff and Hartley 2001c).

Employment

Graduates with specialist degrees are often more marketable than generalists (Bruhn 1995). The problem lessens somewhat as employment in higher budget categories (transition to systems development) is pursued, due to natural merging of disciplines as focused technologies advance into broader systems.

Funding

Many of the large research-sponsoring organizations are structured along the lines of mono-discipline university departments. Their review panels tend to have similar structures, and have the same problems for multi/ inter-disciplinary research as noted above under Evaluation (Bruhn 1995, Butler 1998, Metzger and Zare 1999). In general, mono-discipline research proposals fare better than inter-disciplinary research proposals, except where programs have been specifically designed to fund inter-disciplinary research proposals.

Promotion

The reward system in universities is designed to recognize the research and scholarly contributions of individuals, not teams (Bruhn 1995, Metzger and Zare 1999). Tenure in universities is dependent on the number and quality of publications, and is helped by funds that researchers can attract. As shown above, publications and funding are easier to obtain in mono-disciplinary research, and therefore inter-disciplinary research is penalized further.

Recognition

National academies and other prestigious professional organizations and awards are almost wholly discipline-structured (Metzger and Zare 1999). Since recognition has some dependence on publications and citations, and in many cases on research empires established (funding obtained), mono-disciplinary advantages noted above for publications and funding flow into recognition as well.

BOX 2 – LITERATURE-BASED DISCOVERY

Literature-based discovery surfaces innovative concepts from directly or indirectly-linked literatures. In published or ongoing studies, the following generic steps are used.

1) A problem or topic is defined, and the literature related to that topic is retrieved. For example, in Swanson's most recently published literature-based discovery study, bio-warfare agent prediction was the topic of interest, and a bio-warfare agent literature was retrieved.

2) Then, literatures directly and indirectly related to the initial literature are identified. For example, Swanson's bio-warfare agent study shows that directly related literatures such as virus pathogenicity and virus transmissibility can be identified.

Finally, innovative concepts that exist in these directly and indirectly related literatures, but not in the initial literature, can be identified as true discovery. For example, in Swanson's bio-warfare study, fourteen viruses were hypothesized as potential bio-warfare agents. These viruses have not been tested to confirm Swanson's hypothesis. However, Swanson's approach also identified 17 viruses out of the 21 in Geissler's reference on biological weapons (Geissler 1986), to high statistical significance, offering some confidence in the validity of his predictions on potential bio-warfare agents.

Section 6. Detection of Unexpected Asymmetries from the Biomedical Literature

(based on Kostoff, R. N. "Bilateral Asymmetry Prediction". Medical Hypotheses. 61:2. 265-266. August 2003.)

OVERVIEW

This section outlines the process for detecting unexpected asymmetries in the biomedical literature, and predicts asymmetries in lateral organ cancer incidence from text mining of the Medline database. Lung, kidney, teste, and ovary cancers were examined. For each cancer, Medline case report articles focused solely on 1) cancer of the right organ and 2) cancer of the left organ were retrieved. The ratio of right organ to left organ articles was compared to actual patient incidence data obtained from the National Cancer Institute's (NCI) SEER database for the period 1979-1998. The agreement between the Medline record ratios and the NCI's patient incidence data ratios ranged from within three percent for lung cancer to within one percent for teste and ovary cancer. This is the first known study to generate cancer lateral incidence asymmetries from the Medline database. The technique should be applicable to other diseases and other types of system asymmetries.

BACKGROUND

Unexpected asymmetries in biological or physical systems often stimulate further investigation, with the goal of providing insight into the asymmetry's causes. While the conduct and analysis of laboratory, clinical, or field tests can surface such asymmetries, these approaches can be expensive in labor, capital, and time expenditures. Identification of alternative approaches for surfacing asymmetries could provide an inexpensive option for obtaining this data.

One largely unexplored source of discovery is semi-automated text analysis of large technical literature databases. The author has been exploring different text mining (1,2) approaches to generating innovation and discovery from the large technical literature databases (3,4), and hypothesized that computational linguistics could be used to identify unexpected asymmetries, rapidly and inexpensively. This section confirms that hypothesis.

OBJECTIVES

The purpose of the present study is to ascertain whether asymmetries in lateral organ cancer incidence could be predicted accurately by application of computational linguistics to the Medline database.

APPROACH

Four types of cancers were examined: lung, kidney, teste, ovary. For each cancer, Medline case report articles focused solely on 1) cancer of the right organ and 2) cancer of the left organ were retrieved, using information retrieval techniques (5) developed by the author. For example, to obtain the Medline records focused on cancer of the left kidney, the following query was used: (LEFT KIDNEY OR LEFT RENAL) AND KIDNEY NEOPLASMS AND CASE REPORT[MH] NOT (RIGHT KIDNEY OR RIGHT RENAL). The ratio of numbers of right organ to left organ articles was compared to actual patient incidence data obtained from the NCI's SEER database for the period 1979-1998.

RESULTS

The results are presented in Table 1. The first column contains the organ in which the lateral asymmetry is studied, the second column contains the ratio of Medline case report records focused solely on right organ cancer to those focused solely on left organ cancer, and the third column contains a similar ratio obtained from the NCI SEER database of patient incidence records.

TABLE 1 – RATIO OF RIGHT TO LEFT ORGAN CANCER INCIDENCE

ORGAN	RNK	NCI
LUNG	**1.358**	**1.395**
KIDNEY	**1.024**	**1.043**
TESTE	**1.128**	**1.134**
OVARY	**1.034**	**1.038**

The agreement between the Medline record ratios and the NCI's patient incidence data ratios ranged from within three percent for lung cancer to within one percent for teste and ovary cancer.

CONCLUSIONS AND DISCUSSION

This is the first known study to generate cancer lateral incidence asymmetries from the Medline database. A previous study (6) reported obtaining such ratios by analyzing the ratios of the phrases 'right' and 'left' from patient diagnostic records, although the reason for using the secondary ratio of right/ left phrase frequencies rather than the primary ratio of right/ left record frequencies (i.e., ratio of actual number of patient occurrences) is unclear.

The present study results are based on the assumption that, in a large population, the number of cancer lateral incidence papers published in Medline is proportional to the actual number of cancer lateral incidence occurrences. The excellent agreement of the predictive model with the NCI data provides a strong measure of credibility to this assumption. Such an assumption implies that medical research reported on these cancers treats laterality as a random variable.

Three obvious research questions are suggested by the present study's results.

1. Can the use of text mining of large databases for identification of unexpected asymmetries be extrapolated to other cancers, non-cancer chronic diseases, and other types of systems (biological, physical, environmental, engineering)? Such asymmetries have been identified by the present predictive model for non-cancer chronic diseases (e.g., tuberculosis). A polling of numerous medical experts did not identify any database that contains patient lateral non-cancer chronic disease incidence occurrence, against which the predictive model's results could be compared.

2. Can more advanced text mining (1, 3) be used to provide insights to the mechanisms underlying the asymmetries, serving as another avenue to identifying potential treatments or causes? Additionally, the use of clustering would allow higher resolution of the database's principal themes, and would allow identification of asymmetries at very fine levels of detail.

3. Would further refinement of the query according to the principles contained in (5) reduce the differences between the Medline record ratios and the SEER patient incidence data?

Answers to these research questions are being pursued in parallel. In particular, the fundamental asymmetry detection approach is being applied to identify specific research investment strategy differences among countries, by identifying technical phrase frequency differences in country technology databases. The results are very illuminating, and validate the applicability of the fundamental technique to very diverse problems in disparate disciplines.

REFERENCES FOR SECTION 6.

1. Kostoff, RN, DeMarco, R: Science and technology text mining. Analytical Chemistry. 73:13. 370-378A. 1 July 2001.

2. Kostoff, RN: Text mining for global technology watch. Encyclopedia of Library and Information Science. Second Edition. M. Drake (ed.). Marcel Dekker, Inc. New York, NY. Vol. 4. 2789-2799. 2003.

3. Kostoff, RN: Science and technology innovation. Technovation. 19:10. 593-604. October 1999.

4. Kostoff, RN: Stimulating innovation. International Handbook of Innovation. Chapter 28. Larisa V. Shavinina (ed.). Elsevier Social and Behavioral Sciences. Oxford, UK. 388-400. 2003.

5. Kostoff, RN: The extraction of useful information from the biomedical literature. Academic Medicine. 76:12. 1265-1270. December 2001.

6. Goldman, JA, Chu, WW, Parker, DS, Goldman, RM: Term domain distribution analysis: a data mining tool for text databases. Methods of Information in Medicine. 38: 96-101. 1999.

Section 7. Factor Matrix Text Filtering and Clustering.

(based on Kostoff, R. N., and Block, J. A. "Factor Matrix Text Filtering and Clustering." JASIST. 56:9. 946-968. July. 2005.)

OVERVIEW

The presence of trivial words in text databases can impact record or concept (words/ phrases) clustering adversely. Additionally, the determination of whether a word/ phrase is trivial is context-dependent. The objective of the present section is to demonstrate a context-dependent trivial word filter to improve clustering quality. Factor analysis was used as a context-dependent trivial word filter for subsequent term clustering. Medline records for Raynaud's Phenomenon were used as the database, and words were extracted from the record Abstracts. A factor matrix of these words was generated, and the words that had low factor loadings across all factors were identified, and eliminated. The remaining words, which had high factor loading values for at least one factor and therefore were influential in determining the theme of that factor, were input to the clustering algorithm. Both quantitative and qualitative analyses were used to show that factor matrix filtering leads to higher quality clusters and subsequent taxonomies.

INTRODUCTION

Science and technology (S&T) form the core of modern economies and militaries. Global S&T expenditures are in the neighborhood of 500 billion dollars to a trillion dollars annually, depending on one's definition of S&T. No single organization, or even nation, can begin to cover the full spectrum of S&T development required for a modern competitive economy or military. Cooperative S&T development efforts, leveraging, exploiting, and awareness of external S&T efforts are required if an organization or nation is to remain competitive.

Governments and industrial organizations need ready access to the results of all global research performed in order to:

1) Track research impacts, to help identify benefits arising from sponsored research;
2) Evaluate science and technology programs;
3) Avoid research duplication;
4) Identify promising research directions and opportunities;

5) Perform myriad oversight tasks; and, in general,
6) Support every step of a strategic research planning/ selection/ management/ evaluation process that makes optimal use of S&T investment resources.

In addition, recent counter-terrorism concerns have highlighted the need for ready access to, and analysis of, databases that could link people with institutions and activities. In the S&T arena, this requires linking research performers with organizations, countries, and technical areas.

Complementing this massive global S&T expenditure is equally massive documentation that can be collectively called the global S&T literature. It consists of myriad S&T planning and vision/ requirements documents (S&T planning literature), S&T program descriptive documents (ongoing S&T program literature), evaluation documents of ongoing and completed S&T programs/ projects (S&T program assessment literature), and myriad S&T output and product documents (S&T output literature, such as papers, patents, etc). In order to be able to extract useful information from this massive literature, semi-automated text analysis processes, known collectively as text mining, are required.

One analytical technique commonly used for achieving most of the objectives listed above is a component of text mining: clustering of related textual objects. Clustering is fundamentally a separation process, and is used in many different disciplines, such as isotope separation in chemistry and physics, and impurity separation in water purification. In text analysis, clustering is intrinsically more complicated than in the physical separation processes, because of the multiple meanings and contextual dependence of words, phrases, and word/ phrase patterns.

Additionally, in the S&T text, the high technical content phrases/ words are imbedded in a much larger sea of low technical content words/ phrases, known collectively as trivial words or stop words. If these context-dependent trivial words/ phrases are retained during the clustering process, they can then become the nucleation centers for clustering rather than the desired high technical content words/ phrases, thereby leading to diffuse and misleading clusters. Removing these context-dependent trivial words/ phrases prior to the clustering process would be a major contributor towards defining the clusters more sharply and accurately.

One candidate method for removing context-dependent trivial words prior to clustering is factor analysis (Kendall, 1956; Kaiser, 1960; Cattell, 1966; Cooper, 1983; Coovert and McNelis, 1988; McArdle, 1990; Jackson, 1991; Yalcin and Amemiya, 2001; Browne, 2001), commonly used to identify the pervasive themes in text databases based on correlations, and subsequently group these themes. Factor analysis tends to provide a better estimate of quantitative relationships among the theme components than cluster analysis, whereas cluster analysis tends to provide a better estimate of the structural relationships among the themes, especially hierarchical clustering methods.

Factor and cluster analysis algorithms have existed for decades, and are well validated. Understood less well are how factor and cluster analysis should operate synergistically, and how best to select the data required for input to these algorithms. This study proposes the use of factor analysis as a context-dependent word filter for cluster analysis.

The factor matrix filtering approach first identifies the high technical content words from raw text using factor analysis, and discards the remainder as trivial. It then uses the high technical content words as input to the clustering algorithm. The paper provides an estimate of the benefit of this approach, using a Raynaud's Phenomenon database at the test bed. This paper starts with the Background of the various study elements, describes the Approach used, presents the Results, and ends with Conclusions.

BACKGROUND

As summarized in the Introduction, this section includes text mining, clustering, trivial word removal, factor analysis, and Raynaud's Phenomenon. The present sub-section will provide sufficient background material on each of these topics to clarify the procedures described in the Approach sub-section.

Text Mining

Text mining has been developed to extract useful information from the global S&T literature, in order to supplement conventional human-based approaches (Hearst, 1999; Trybula, 1999; Feldman, 1999; Lagus et al, 1999; Weiss et al, 1999; Losiewicz et al, 2000; Kuhnhold, 2000; Visa, 2001; Kostoff et al, 2001c; Zhu and Porter, 2002; Kogan et al, 2003; Perrin and

Petry, 2003). Its component capabilities of *computational linguistics* and *bibliometrics* can be summarized as follows.

Science and technology *computational linguistics* [Kostoff, 2003a; Hearst, 1999; Zhu and Porter, 2002; Losiewicz et al, 2000] is a process that underlies the extraction of useful information from large volumes of technical text. It identifies pervasive technical themes in large databases from technical phrases that occur frequently. It also identifies relationships among these themes by grouping (clustering) these phrases (or their parent documents) on the basis of similarity. *Computational linguistics* can be used for:

- Enhancing information retrieval and increasing awareness of the global technical literature [Kostoff et al, 1997; Greengrass, 1997; TREC, 2003]
- Potential discovery and innovation based on merging common linkages among very disparate literatures [Swanson, 1986; Swanson and Smalheiser, 1997; Kostoff, 2003b; Gordon and Dumais, 1998]
- Uncovering unexpected asymmetries from the technical literature [Goldman et al, 1999; Kostoff, 2003c]
- Estimating global levels of effort in S&T sub-disciplines [Kostoff et al, 2000a, 2002, 2004a; Viator and Pestorius, 2001]
- Helping authors potentially increase their citation statistics by improving access to their published papers, and thereby potentially helping journals to increase their Impact Factors [Kostoff et al, 2004a, 2004b]
- Tracking myriad research impacts across time and applications areas [Davidse and VanRaan, 1997; Kostoff et al, 2001b].

Evaluative *bibliometrics* [Narin, 1976; Garfield, 1985; Schubert et al, 1987] uses counts of publications, patents, citations and other potentially informative items to develop science and technology performance indicators. Its validity is based on the premises that 1) counts of patents and papers provide valid indicators of R&D activity in the subject areas of those patents or papers, 2) the number of times those patents or papers are cited in subsequent patents or papers provides valid indicators of the impact or importance of the cited patents and papers, and 3) the citations from papers to papers, from patents to patents and from patents to papers provide indicators of intellectual linkages between the organizations that are producing the patents and papers, and knowledge linkage between their subject areas [Narin et al, 1994]. Evaluative *bibliometrics* can be used to:

- Identify the infrastructure (authors, journals, institutions) of a technical domain (Kostoff et al, 2000b, 2004b),
- Identify experts for innovation-enhancing technical workshops and review panels,
- Develop site visitation strategies for assessment of prolific organizations globally,
- Identify impacts (literature citations) of individuals, research units, organizations, and countries (Kostoff et al, 2001b, 2004c)

Evaluative bibliometrics can also used to help generate extensive background material for research papers, and for comprehensive literature reviews and surveys. The documents most cited (relative to their contemporaries) by a retrieved topical literature of interest can be considered to be seminal, and form the core of the background material. Other relevant documents can be added to enhance the background material and eliminate gaps in the narration. Another advantage of this citation-assisted background (CAB) approach (Kostoff, 2004d) over traditional literature reviews is that the core seminal papers identified are based on the larger technical community's consensus (highest citations), rather than solely based on the author(s) personal experiences and biases.

A typical text mining study of the published literature develops a query for comprehensive information retrieval, processes the retrieved database using computational linguistics and bibliometrics, and integrates the processed information.

Clustering

Clustering is grouping by attributes. Since technical text can contain many attributes, many taxonomies can be generated from a body of text. Two types of clustering that have been used extensively by the first author are concept clustering and document clustering, Concept clustering is the grouping of related words or phrases to identify technical themes in the text database. It has been used to generate literature taxonomies (Kostoff and DeMarco, 2001c), to facilitate Web searching (Khare, 2003), to summarize text (Ko et al, 2003), to generate hypotheses and discovery (Stegmann and Grohmann, (2003), and to generate thesauri (Hodge and Austin, 2002). Document clustering (Cutting et al, 1992a; Guha et al, 1998; Hearst et al, 1998; Karypis et al, 1999; Rasmussen, 1992; Steinbach et al, 2000; Willet,

1988; Zamir and Etzioni, 1998; Karypis, 2002; Guerrero-Bote et al, 2003; Casillas et al, 2003; Schenker et al, 2003) is the grouping of related documents by theme.

Clusters can be aggregated into a hierarchical structure, to provide a taxonomy or classification scheme of the discipline(s) being studied. The quality of the final clusters and taxonomy is strongly dependent on the quality of the words selected for input to the factor and cluster analyses. If important high technical content words and phrases are omitted from the input, the themes derived from these words will be lost to the final results. If too many non-technical words are selected for the input, then artificial clusters will be generated based on overlap of non-technical words, and/ or words/ phrases will be re-assigned among clusters due to non-technical linkages. A misleading taxonomy will result.

One of the reviewers of this section asked how the proposed improved clustering approach could lead to improved information retrieval from Medline (the source of the Raynaud's records used as the database). The response depends on how clustering is employed in the information retrieval process.

In the first author's Simulated Nucleation information retrieval approach (Kostoff et al, 1997), clustering (including the factor matrix clustering approach presented in the present paper) is used to reduce time spent on query development, as well as to expand the query. This iterative relevance feedback query expansion technique starts with generation of a test query. Records are then retrieved from a source database (e.g., Medline) using this query. The next series of steps involves separating the relevant from non-relevant records, and identifying text patterns characteristic of relevant records but essentially non-existent in non-relevant records, and vice versa. The query is then modified with these patterns so that it will retrieve more relevant records, and will filter out more non-relevant records.

Before clustering was used by the first author to supplement the information retrieval process, the separation of relevant from non-relevant records was performed manually. Many records (Abstracts, if records are journal articles) were read, and the relevant and non-relevant were separated. Computational linguistics were performed on each group (relevant, non-relevant) of records, to identify the text patterns characteristic of each group.

For the past few years, the first author has used clustering for the relevant/ non-relevant separation. Records are retrieved, then immediately clustered. In almost all cases, the relevant records will cluster together, as will the non-relevant records. Substantial time is saved by substituting computer-based clustering for manual separation. Additionally, when selecting additional words for the query, words can be selected from all the clusters to insure that all the main concepts have at least some representation in the query. Factor matrix filtering provides sharper clusters, insuring that query terms selected will contribute to improving recall and precision.

Trivial Word Removal

An intrinsic problem in any text grouping procedure, or indeed in any text feature extraction procedure, is dilution of the results by the inclusion of trivial words/ phrases. In technical literature, trivial words/ phrases are terms with little technical information content. The most common trivial words are generic connectors such as 'the', 'of', 'and', etc. They tend to have low technical content regardless of context. Many other words become trivial in selected contexts.

The presence of trivial words can alter the results of clustering substantially. Experiments by the first author on the details of technical text clustering in project narratives and journal article Abstracts showed that, on average, there were very few non-trivial words in each document that were important for determining the main theme of the article. In an ideal world, these few words would serve as the anchors for clustering, and produce sharp well-defined clusters. Unfortunately, the presence of the trivial words/ phrases provides competing anchors for the clusters, and generates clusters and taxonomy structures that are less-well defined. As stated in (Wang et al, 2003), "In many document data sets, only a relatively small number of the total features may be useful in classifying documents, and using all the features may adversely affect performance". Any techniques that would pre-process data prior to clustering, and eliminate any words/ phrases that are trivial in the context of the application, would help sharpen the resultant clusters.

Most, if not all, text mining approaches start this word selection/ filtering process by attempting to reduce the dimensionality of the system (reducing the number of words or phrases to be manipulated). After the raw text of the literature to be analyzed is converted to words or phrases, all text mining

approaches separate out the ‘trivial’ terms from the high technical content terms. Almost all the approaches reported in the literature assume the ‘trivial’ terms are context-independent, and these approaches use a pre-determined trivial word or stop list to remove such words from the initial pool of words.

One innovative approach for identifying context-dependent trivial words selected articles about genes from diverse classes, then eliminated the high frequency words in the total dataset by assuming these high frequency words were generic rather than gene specific, and would serve to diffuse the clustering process (Kankar et al, 2002).

Some approaches use statistical techniques to eliminate trivial words/ phrases. A key concept is that words/ phrases that appear uniformly over the collection are trivial from the viewpoint of theme discrimination among documents/ concepts. Weighting terms based on maximal tf*idf (term frequency-inverse document frequency) is a popularly-used filter (Salton and Buckley, 1998). Feldman et al (1998) use the standard deviation of the relative frequency of a term over all the documents of the collection as a weighting. Stensmo’s probabilistic information retrieval system (Stensmo, 2002) bypasses the need for stop-word lists by removing unneeded parameters dynamically based on a local mutual information measure.

Perhaps the most extensive work in this area is attributable to Bookstein (Bookstein et al, 1995, 1998, 2003) and Wilbur (Wilbur and Sorotkin, 1992; Kim and Wilbur, 2001; Wilbur and Yang, 1996). Bookstein asserts that the greater the deviation of a term’s distribution from a Poisson distribution, the more likely that the term is a useful one. His approach examines term clumping tendencies in full text, and is not very relevant to Abstracts or titles. Wilbur examines a variety of statistical measures to estimate term importance, including term strength (how strongly the term's occurrences correlate with the subjects of documents in the database).

Other techniques use noun phrase extraction based on natural language processing, such as part-of-speech tagging and filtering (Brill, 1993, 1994; Cutting, 1992b). Identification of parts of speech may not be easily made for literature that describes the latest science and technology, with its continual infusion of new terminology. Lexicons, dictionaries, and thesauri have an intrinsic lag time, and consequently new terms may appear

incomprehensible to any computer-based tagger. Additionally, there are both semantic and syntactic components to identifying high quality phrases. While tagging can address the syntactic issue, it cannot eliminate the semantic problem. As stated in the above papers, all the statistical feature selection approaches and part-of-speech tagging approaches leave much to be desired, mainly because of the influence of context on feature desirability.

After the words/ phrases have been selected initially, almost all techniques use some further processing of the words/ phrases before input to the factor or clustering algorithms. Many techniques use stemming (e.g., Porter, 1980; Kostoff, 2003d), where words are grouped according to common roots, and conflated (singulars are combined with their plurals, full spellings are combined with their acronyms, and different tenses are combined). A few techniques select synonyms from a controlled vocabulary to reduce the number of words while retaining the concepts (see Weeber et al (2001), which used a medical thesaurus for this purpose in literature-based discovery). Finally, for those techniques that use both factor analysis and clustering, each approach is pursued independently.

It is the contention of the authors that the words to be selected as input to the cluster analyses should be context-dependent. 'High-technical content' has different meanings for different literatures and applications. 'Trivial' has different levels of context dependency. Stemming and conflation should be dependent on context as well. This section will show how context-dependency can be used in the word or phrase selection process through factor matrix filtering. A companion paper shows how context-dependency can be incorporated in the conflation process through factor matrix filtering (Kostoff, 2003d).

Factor Analysis

Factor analysis of a text database aims to reduce the number of words/ phrases (variables) in a system, and to detect structure in the relationships among words/ phrases. Word/ phrase correlations are computed, and highly correlated groups (factors) are identified. The relationships of these words/ phrases to the resultant factors are displayed clearly in the factor matrix, whose rows are words/ phrases and columns are factors. In the factor matrix, the matrix elements Mij are the factor loadings, or the contribution of word/ phrase i (in row i) to the theme of factor j (in column j). The theme of each factor is determined by those words/ phrases that have the largest values of

factor loading. Each factor has a positive value tail and negative value tail. For each factor, one of the tails dominates in terms of absolute value magnitude. This dominant tail is used to determine the central theme of each factor.

One of the key challenges in factor analysis is defining the number of factors to select. Different approaches have been suggested in the literature, but the two most widely used are the Kaiser criterion (Kaiser, 1960; Jackson, 1991), and the Scree test (Cattell, 1966). The Kaiser criterion states that only factors with eigenvalues greater than unity should be retained, essentially requiring that a factor extracts at least as much variance as the equivalent of one original variable. The Scree test plots factor eigenvalue (variance) vs factor number, and recommends that only those factors that extract substantive variance be retained. Operationally, the factor selection termination point becomes the 'elbow' of the Scree plot, the point where the slope changes from large to small.

As one of the reviewers of this section correctly noted, the "interpretation of the Scree Plot is partly subjective", and "consistency among different interpreters is low". Appendix 1 discusses this interpretation problem in more detail, and shows that part of the inconsistency may stem from the fractal-like nature of the Scree Plots. Appendix 1 also discusses the consequences of selecting factor matrices with different numbers of factors.

In most previous studies performed by the first author, the Kaiser criterion has been used to select the number of factors for the factor matrix. These previous studies have used an Excel add-in to generate the factor matrices, and, due to Excel's limitations on columns, have been limited approximately to 250 x 250 correlation matrices, or 250 words. The Kaiser criterion has yielded factor numbers in the range of 20-45, considered a reasonable number for analysis. However, in the present Raynaud's Phenomenon study, another software package that did not require Excel was used (TechOasis), and 659 words were used for the correlation matrix. The Kaiser criterion yielded 224 factors, a number far too large for detailed factor analysis, and of questionable utility, since many of the eigenvalues were not too different from unity. It was decided to examine the Scree Plot for factor number determination.

Raynaud's Phenomenon

Both authors are conducting a study of Raynaud's Phenomenon (a peripheral circulatory disorder) using text mining, to extend the literature-based discovery techniques from Swanson's classical paper (Swanson, 1986). Of central interest are the pervasive medical themes of the Raynaud's Phenomenon literature, identified by factor analysis and cluster analysis.

Since each factor from the factor analysis, or cluster from the cluster analysis, addresses some aspect of Raynaud's Phenomenon, an overview of Raynaud's Phenomenon will be presented before discussing the factor and cluster results. Because the main Raynaud's terminology used in the literature is not consistent (in many cases, Raynaud's Disease is used interchangeably with Raynaud's Phenomenon or Raynaud's Syndrome), the overview will include the distinction among these Raynaud variants.

Raynaud's Phenomenon is a condition in which small arteries and arterioles, most commonly in the fingers and toes, go into spasm (contract) and cause the skin to turn pale (blanching) or a patchy red (rubor) to blue (cyanosis). While this sequence is normally precipitated by exposure to cold, and subsequent re-warming, it can also be induced by anxiety or stress. Blanching represents the ischemic (lack of adequate blood flow) phase, caused by digital artery vasospasm. Cyanosis results from de-oxygenated blood in capillaries and venules (small veins). Upon re-warming, a hyperemic phase ensues, causing the digits to appear red.

Raynaud's Phenomenon can be a primary or secondary disorder. When the signs of Raynaud's Phenomenon appear alone without any apparent underlying medical condition, it is called Primary Raynaud's, or formerly, Raynaud's Disease. In this condition, the blood vessels return to normal after each episode. Conversely, when Raynaud's Phenomenon occurs in association with an underlying condition or is due to an identifiable cause, then it is referred to as Secondary Raynaud's, or formerly, as Raynaud's Syndrome. The most common underlying disorders associated with Secondary Raynaud's are the auto-immune disorders, or conditions in which a person produces antibodies against his or her own tissues. In contrast to Primary Raynaud's, where the blood vessels remain anatomically normal after each episode, in Secondary Raynaud's there may be scarring and long-term damage to the blood vessels; thus Secondary Raynaud's is potentially a more serious disorder than Primary. Certain repetitive activities may result in a predisposition to Raynaud's Phenomenon. These cases of so-called

"Occupational Raynaud's" typically result from the chronic use of vibrating hand tools.

Thus, while Raynaud's Phenomenon is a direct consequence of reduced blood flow due to reversible blood vessel constriction, it may be a function of many variables that can impact blood flow. These include:

*Inflammation from the auto-immune disorders that can cause swelling and thereby constrict blood vessels;
*Increased sympathetic nervous system activity, that can affect the timing and duration of the blood vessel muscular contractions that cause constriction;
*Heightened digital vascular reactivity to vaso-constrictive stimuli, that cause the blood vessels to over-react and over-contract;
*Deposits along the blood vessel walls that can reduce blood flow and increase the flow sensitivity to contraction stimuli;
*Blood rheological properties that offer additional resistance to blood flow, and magnify the impact of blood vessel constriction;
*Blood constituents and hormones that can act as vaso-constrictors or vaso-dilators.

APPROACH

In the first part of this study, 930 Medline Abstract-containing records related to Raynaud's Phenomenon, and published in the 1975-1985 time period (to approximate Swanson's database), were retrieved with a Raynaud's-specific query. These Abstracts were subjected to factor analysis and clustering, as part of the analysis.

All the single words were automatically extracted from the database of 930 Abstracts, subject to elimination of very trivial words (removing words like 'of', 'the', 'and', 'if', etc). Non-trivial single words (659) were then manually extracted (by experts) from the database of Abstracts, along with the number of documents in which each word appeared (document frequency). For this database, the 659 extracted words were near the limit that allowed a factor matrix to be computed. The co-occurrence of word pairs in the same document (word co-occurrence frequency) was computed, and a correlation matrix (659 x 659) of word pairs was generated. The variables were factorized, and a factor matrix was generated.

Factor Matrix Generation

Once the desired number of factors has been determined from the Scree Plot 'elbow', and the appropriate factor matrix has been generated, the factor matrix can then be used as a filter to identify the significant technical words for further analysis. Specifically, the factor matrix can complement a basic trivial word list (i.e., a list containing words that are trivial in almost all contexts, such as 'a', 'the', 'of', 'and', 'or', etc) to select context-dependent high technical content words for input to a clustering algorithm. The factor matrix pre-filtering will improve the cohesiveness of clustering by eliminating those words that are trivial words operationally in the application context.

The variance accounted for by each underlying factor (eigenvalue) was generated by Principal Components Analysis. Figure 1 shows the factor eigenvalue-factor number plot (Scree Plot) for the 659 un-rotated factors on a linear scale. The 'elbow', or break point, of the curve appears to be about fourteen factors.

INSERT FIGURE 1

Factor Matrix Filtering

The fourteen factor matrix determined by the Scree Plot of Figure 1 was examined in detail, and is presented in the Results section. To diversify the factor loading patterns, and simplify interpretation of each factor, varimax orthogonal rotation was used.

Factor Matrix Word Filtering and Selection

After the factor matrix has been generated, its highest technical content words are input to the clustering algorithm. In the present experiment, the 659 words in the factor matrix would have to be culled to the ~250 allowed by the Excel-based clustering package, WINSTAT. The ~250 word limit is an artifact of Excel. Other software packages may allow more or less words to be used for clustering, but all approaches perform culling to reduce dimensionality. The filtering process presented here is applicable to any level of filtered words desired.

Another caveat. A trivial word list of the type described previously (words that are trivial in almost all contexts) was used to arrive at the 659 words used for the factor matrix input. This was not necessary. The raw words from the word generator could be used as input, and would be subject to the same filtering process. To allow more important words to be used in this demonstration, the very trivial words were removed.

The factor loadings in the factor matrix were converted to absolute values. Then, a simple algorithm was used to automatically extract those high factor loading words at the dominant tail of each factor. The highest absolute value of factor loading for each word/ phrase was identified from the total factor matrix. The words/ phrases were ranked in inverse order of highest absolute value of factor loading. If word variants were on this list (e.g., singles and plurals), and their factor loadings were reasonably close (Kostoff, 2003d), they were conflated (e.g., 'agent' and 'agents' were conflated into 'agents', and their frequencies were added). All words/ phrases below the ~250 term limit allowed by Excel were eliminated.

RESULTS

Factor Matrix Analysis

For the fourteen factor matrix, the high factor loading words in the dominant tail of each factor are shown in parentheses after the factor number, followed by a brief narrative of the factor theme.

Factor 1 (nuclear, antibodies, extractable, speckled, connective, immuno-fluorescence, antinuclear, tissue, anti-RNP, MCTD, mixed, ribonucleoprotein, swollen, RNP, antibody, antigen, titer, SLE, lupus, erythematosus) focuses on different types of autoantibodies, especially anti-nuclear and extractable nuclear, and their relation to auto-immune diseases.

Factor 2 (double-blind, placebo, mg, daily, weeks, times, agent, nifedipine, trial) focuses on double-blind trials for vasodilators.

Factor 3 (vibration, tools, workers, vibrating, exposure, chain, prevalence, time, exposed, sensory, white, circulatory, complaints) focuses on the impact of vibratory tools on circulation.

Factor 4 (coronary, ventricular, heart, angina, hypertension, myocardial, cardiac, failure, pulmonary) focuses on coronary circulation and blood pressure problems.

Factor 5 (prostaglandin, platelet, E1, prostacyclin, aggregation, infusion, hours, healing, ischaemic, thromboxane, administered, vasodilator, intravenous) focuses on the administration of vasodilators to improve circulation.

Factor 6 (calcinosis, sclerodactyly, esophageal, dysmotility, telangiectasia, anticentromere, variant, diffuse, scleroderma) focuses on scleroderma-spectrum types of autoimmune diseases.

Factor 7 (extremity, sympathectomy, artery, surgery, arteries, upper, occlusions, arterial, brachial, thoracic, operation, surgical, angiography, occlusive) focuses on surgical solutions to remove constrictions on circulation.

Factor 8 (C, degrees, systolic, pressure, cooling, blood, finger, measured, flow) focuses on blood flow, and associated finger blood pressure and temperature measurements.

Factor 9 (capillaries, capillary, nail-fold, microscopy, capillaroscopy) focuses on the diagnostic use of nail-fold capillary microscopy.

Factor 10 (training, biofeedback, relaxation, stress, outcome, measures, headaches, temperature, conducted, thermal, physiological, responses) focuses on the use of biofeedback training to reduce stress headaches, and raise temperatures through improved circulation.

Factor 11 (vasodilation, peripheral, immersion, calcium, water) focuses on vasodilation of the peripheral circulatory system after immersion, and the role of calcium in this process.

Factor 12 (complexes, immune, circulating, complement, IgG, serum, levels, IgM) focuses on serum levels of circulating immune complexes and immunoglobulins, especially IgG and IgM.

Factor 13 (eosinophilia, fasciitis, fascia, eosinophilic, visceral, hypergammaglobulinemia, absent, scleroderma-like, corticosteroids) focuses on inflammation, especially of the fascia.

Factor 14 (systemic, lupus, RA, erythematosus, PSS, sclerosis, rheumatoid, arthritis, SLE) focuses on autoimmune diseases associated with Raynaud's Phenomenon.

The fourteen factor matrix themes can be divided into the two main thrusts of circulation and autoimmunity, where circulation covers factors 2, 3, 4, 5, 7, 8, 9, 10, and 11, and autoimmunity covers factors 1, 6, 12, 13, 14.

An examination of the words eliminated and those retained showed that most of those retained appeared to have high technical content, and would have been selected by previous manual filtering processes for input to the clustering algorithms. Some of the words in isolation appeared not to have the highest technical content, also as shown above, but it was concluded that they were important because of their contribution to theme determination in the present clustering application. Similarly, some of the words eliminated by the factor matrix filter appeared to be high technical content, and in previous manual filtering processes might have been selected for the clustering algorithm input (e.g., acrocyanosis, vasomotor, cerebral, gastrointestinal). The conclusion for these words was not that they were unimportant per se. Rather, they did not have sufficient influence in determining the factor themes, and would not make an important contribution to the cluster structure determination. Thus, the context dependency (their influence on factor theme determination) of the words was the deciding factor in their selection or elimination, not only the judgement of their technical value in isolation (independent of factor theme determination), as was done in previous manual filtering approaches.

Word Clustering

The 252 filtered and conflated words were input to the WINSTAT clustering algorithm, and the Average Link option was selected for clustering. Figure 2 is the dendrogram of the 252 words. This is a tree-like structure that shows how the individual words cluster into groups in a hierarchical structure. One axis is the words, and the other axis ('distance') reflects their similarity. The lower the value of 'distance' at which words, or word groups, are linked together, the closer their relation. As an extreme case of

illustration, words that tend to appear as members of multi-word phrases, such as ‘lupus erythematosus’, ‘connective tissue’, or ‘double blind’ appear adjacent on the dendrogram with very low values of ‘distance’ at their juncture.

INSERT FIGURE 2

Now the major structures, or clusters, will be described, following the hierarchical structure of the dendrogram. The capitalized words listed in parentheses after each cluster number are the boundaries of that cluster from the dendrogram. Only the top three hierarchical levels will be described.

The top hierarchical level can be divided into two major clusters. Cluster 1 (PATIENTS-OLD) focuses on autoimmunity, and Cluster 2 (TREATMENT-ACID) focuses on circulation. The second hierarchical level can be divided into four clusters, where Cluster 1 is divided into Clusters 1a and 1b, and Cluster 2 is divided into Clusters 2a and 2b. Cluster 1a (PATIENTS-NEUROPATHY) focuses on autoimmune diseases and antibodies, while Cluster 1b (LESIONS-OLD) focuses on inflammation, especially fascial inflammation. Cluster 2a (TREATMENT-CONSECUTIVE) focuses on peripheral vascular circulation, while Cluster 2b (PULMONARY-ACID) focuses on coronary vascular circulation. These four high level categories are not computer artifacts, but correspond extremely well to how medical problems with a Raynaud’s Phenomenon component are diagnosed and treated in medical practice.

Most of the clusters in the second hierarchical level can be rationally divided into two sub-clusters, to produce the third hierarchical level clusters. Cluster 1a1 (PATIENTS-MARKER) has multiple themes: different types of antibodies, especially anti-nuclear and extractable nuclear, and their relation to autoimmune diseases; sclerotic types of autoimmune diseases; and autoimmune diseases associated with Raynaud’s Phenomenon. It incorporates the themes of Factors 1, 6, and 14. Cluster 1a2 (SERUM-NEUROPATHY) focuses on circulating immune complexes, and parallels the theme of Factor 12. Cluster 1b (LESIONS-OLD) is too small to subdivide further, and stops at the second hierarchical level. It parallels the theme of Factor 13.

Cluster 2a1 (TREATMENT-RESERPINE) has multiple themes: double-blind clinical trials for vasodilators; administration of vasodilators to reduce platelet aggregation and improve circulation; blood flow, and associated finger blood pressure and temperature measurements; and occupational exposures, mainly vibrating tools and vinyl chloride, that impact the peripheral and central nervous systems and impact circulation. It incorporates the themes of Factors 2, 3, 5, 7, 8. Cluster 2a2 (CAPILLARY-CONSECUTIVE) focuses on nailfold capillary microscopy as a diagnostic for for microcirculation, and parallels the theme of Factor 9. Cluster 2b1 (PULMONARY-LUNG) focuses on cardiovascular system problems, and parallels the theme of Factor 4. Cluster 2b2 (BIOFEEDBACK-ACID) focuses on biofeedback training to reduce stress and headaches, and increase relaxation, and parallels the theme of Factor 10.

Thus, use of the factor matrix for context-dependent trivial word elimination has produced a taxonomy that is technically defensible. What is the evidence that this taxonomy is improved compared to a taxonomy resulting from non-use of factor matrix filtering? There are two tandem approaches for comparing the quality of taxonomies, quantitative and qualitative. The quantitative approach identifies metrics for gauging the cohesiveness and uniqueness of clusters, and evaluates the performance of each taxonomy against the metrics. The qualitative approach examines the overall taxonomy structure as well as the individual clusters for reasonableness and technical defensibility.

Appendix 2 describes the marginal impact on taxonomy quality from the substitution of trivial words for technical content words. Appendix 2 shows that, for the word clustering approach of this section, the substitution of trivial words for higher technical content words has both quantitative and qualitative consequences. First, the ‘distance’ of the overall dendrogram (the ‘distance’ of the final aggregation-highest point on the dendrogram) increases when the trivial words are inserted. The effect of adding trivial words is to reduce the sharpness and uniqueness of individual clusters, and enhance linkages among disparate clusters through the trivial words only. Since ‘distances’ are low when words in a cluster are very closely related, ‘distances’ increase as the relations become more diffuse.

Second, the trivial words can act as a magnet, and change the balance of words in a cluster. In Appendix 2, the two trivial words used for replacement (the highest frequency words ‘of’, ‘the’) appeared on the

dendrogram in the same first level cluster. They, in turn, attracted words formerly in the other first level cluster, and in some cases, these words were shifted to the less defensible technical category.

Appendix 2 shows further that use of the factor matrix filtering for selecting the words input to the cluster above (compared to use of the highest frequency words with no filtering) had two consequences. The overall taxonomy 'distance' decreased with the use of factor matrix filtering, and the word assignment to clusters improved from a technical perspective. Thus, factor matrix filtering improved the quality of the taxonomy and its constituent clusters. The only negative feature of factor matrix filtering is the modest time required for incorporation of this additional analytical step.

DISCUSSION AND CONCLUSIONS

Factor matrix filtering is an effective method for identifying the major themes in a text database, identifying the critical words that define the theme, selecting these critical words in context for clustering, and identifying which variants of these words can be conflated within the context of the specific database examined.

REFERENCES FOR SECTION 7

Bookstein, A., Klein, S.T., Raita, T. Detecting content-bearing words by serial clustering. (1995). *Proc. 18-th ACM-SIGIR Conf.,* Seattle, WA. 319-327.

Bookstein, A., Klein, S.T., Raita, T. (1998). Clumping properties of content-bearing words. Journal of the American Society for Information Science. 49 (2). 102-114. Feb.

Bookstein, A., Kulyukin, V., Raita, T., Nicholson, J. (2003). Adapting measures of clumping strength to assess term-term similarity. Journal of the American Society for Information Science and Technology. 54 (7). 611-620. May.

Brill, E. (1993). Automatic grammar induction and parsing free text: A transformation-based approach. Proceedings of the 31st meeting of the Association of Computational Linguistics. Columbus, OH.

Brill, E. (1994). Some advances in transformation-based part of speech tagging. Proceedings, Twelfth National Conference on Artificial Intelligence. Seattle, WA.

Browne, M.W. (2001). An overview of analytic rotation in exploratory factor analysis. Multivariate Behavioral Research. 36 (1). 111-150.
Casillas, A, de Lena, M.T.G., and Martinez, R. (2003). Document clustering into an unknown number of clusters using a genetic algorithm. Text, Speech and Dialogue, Proceedings, 2807. Lecture Notes in Artificial Intelligence. 43-49.
Cattell, R.B. (1966). The Scree Test for the number of factors. Multivariate Behavioral Research. 1. 245 –276.
Cooper, J.C.B. (1983). Factor-Analysis - An overview. American Statistician, 37 (2). 141-147.
Coovert, M.D., and McNelis, K. (1988). Determining the number of common factors in factor-analysis - A Review and Program. Educational and Psychological Measurement, 48 (3). 687-692. Fall.
Cutting, D. R., Karger, D. R, Pedersen, J. O. and Tukey, J. W. (1992a). Scatter/Gather: A cluster-based approach to browsing large document collections. In *Proceedings of the 15th International ACM SIGIR Conference on Research and Development in Information Retrieval* (SIGIR'92). 318-329.
Cutting, D., Kupiec, J., Pederson, J., and Sibun, P. (1992b). A practical part of speech tagger. Presented at the Third International ACL Conference on Applied Natural Language Processing. Trento, Italy.
Davidse, R.J., Van Raan, A.F.J. (1997). Out of particles: impact of CERN, DESY, and SLAC research to fields other than physics. Scientometrics 40:2 . 171-193.
Feldman, R., Fresko, M., Kinar, Y., Lindell, Y., Liphstat, O., Rajman, M., Schler, Y., and Zamir, O. (1998). Text mining at the term level. Principles of Data Mining and Knowledge Discovery. Lecture Notes in Artificial Intelligence. 1510. 65-73.
Feldman, R. (1999). Text mining via information extraction. Principles of Data Mining and Knowledge Discovery. 1704. 165-173.
Garfield, E. (1985). History of citation indexes for chemistry - a brief review. JCICS. 25(3): 170-174.
Goldman, J.A., Chu, W.W., Parker, D.S., Goldman, R.M. (1999). Term domain distribution analysis: a data mining tool for text databases. Methods of Information in Medicine. 38. 96-101.
Gordon, M.D., Dumais, S. (1998). Using latent semantic indexing for literature based discovery. Journal of the American Society for Information Science. 49 (8): 674-685.
Greengrass, E. (1997). Information retrieval: An overview. National Security Agency. TR-R52-02-96.

Guerrero-Bote, V.P., Lopez-Pujalte, C., de Moya-Anegon, F., Herrero-Solana, V. (2003). Comparison of neural models for document clustering. International Journal of Approximate Reasoning, 34 (2-3). 287-305. Nov.
Guha, S., Rastogi, R. and Shim, K. (1998). CURE: An efficient clustering algorithm for large databases. In *Proceedings of the ACM-SIGMOD 1998 International Conference on Management of Data* (SIGMOD'98). 73-84.
Hearst, M. A. (1998). The use of categories and clusters in information access interfaces. In T. Strzalkowski (ed.), Natural Language Information Retrieval. Kluwer Academic Publishers.
Hearst, M. A. (1999). Untangling text data mining. Proceedings of ACL 99, the 37th Annual Meeting of the Association for Computational Linguistics. University of Maryland. June 20-26.
Hodge, V.J., and Austin, J. (2002). Hierarchical word clustering - automatic thesaurus generation. Neurocomputing, 48. 819-846. Oct.
Jackson, J. E. (1991). A users guide to principal components. Wiley, New York, NY.
Kaiser, H.F. (1960). The application of electronic computers to factor analysis. Educational and psychological measurement. 20. 141-151.
Kankar, P., Adak, S., Sarkar, A., Murali, K., and Sharma, G. (2002). MedMesh summarizer: Text mining for gene clusters. Proceedings of the Second SIAM International Conference on Data Mining. Robert Grossman, Jiawei Han, Vipin Kumar, Heikki Mannila, and Rajeev Motwani, Editors. April. Arlington, VA.
Karypis, G., Han, E.H., and Kumar, V. (1999). Chameleon: A hierarchical clustering algorithm using dynamic modeling. IEEE Computer: Special Issue on Data Analysis and Mining 32(8). 68-75.
Karypis, G. (2002). CLUTO—A clustering toolkit. http://www.cs.umn.edu/˜cluto.
Kendall, M.G., and Lawley, D.N. (1956). The principles of factor-analysis. Journal of the Royal Statistical Society Series A-General, 119 (1). 83-84.
Khare, A. (2003). Connecting word clusters to represent concepts with application to web searching. Knowledge-Based Intelligent Information and Engineering Systems, Pt 1, Proceedings, 2773: 816-823. Lecture Notes in Artificial Intelligence.
Kim, W, and Wilbur, W.J. (2001). Corpus-based statistical screening for content-bearing terms. Journal of the American Society for Information Science and Technology. 52 (3). 247-259. Feb 1.

Ko, Y., Kim, K., and Seo, J. (2003). Topic keyword identification for text summarization using lexical clustering. IEICE Transactions on Information and Systems, E86D (9): 1695-1701 Sep.
Kogan, J., Nicholas, C., and Volkovich, V. (2003). Text mining with information - theoretic clustering. Computing in Science & Engineering, 5 (6). 52-59. Nov-Dec.
Kostoff, R.N., Eberhart, H.J., and Toothman, D.R. (1997). Database Tomography for information retrieval. Journal of Information Science. 23:4. 301-311.
Kostoff, R.N., Green, K.A., Toothman, D.R., and Humenik, J.A. (2000a). Database Tomography applied to an aircraft science and technology investment strategy. Journal of Aircraft. 37:4. 727-730.
Kostoff, R.N., Braun, T., Schubert, A., Toothman, D.R., and Humenik, J.A. (2000b). Fullerene roadmaps using bibliometrics and Database Tomography. Journal of Chemical Information and Computer Science. 40(1): 19-39.
Kostoff, R.N., and DeMarco, R.A. (2001a). Science and technology text mining. Analytical Chemistry. 73:13. 370-378A. 1 July.
Kostoff, R.N., Del Rio, J.A., García, E.O., Ramírez, A.M., Humenik, J.A. (2001b). Citation mining: integrating text mining and bibliometrics for research user profiling. Journal of the American Society for Information Science and Technology. 52:13. 1148-1156.
Kostoff, R.N., Toothman, D.R., Eberhart, H.J., and Humenik, J.A. (2001c). Text mining using database tomography and bibliometrics: A review. Technology Forecasting and Social Change. 68:3. November.
Kostoff, R.N., Tshiteya, R., Pfeil, K.M., and Humenik, J.A. (2002). Electrochemical power source roadmaps using bibliometrics and database tomography. Journal of Power Sources. 110:1. 163-176.
Kostoff, R.N. (2003a). Text mining for global technology watch. In Encyclopedia of Library and Information Science, Second Edition. Drake, M., Ed. Marcel Dekker, Inc. New York, NY. Vol. 4. 2789-2799.
Kostoff, R.N. (2003b). Stimulating innovation. International Handbook of Innovation. Larisa V. Shavinina (ed.). Elsevier Social and Behavioral Sciences, Oxford, UK.
Kostoff, R.N. (2003c). Bilateral asymmetry prediction. Medical Hypotheses. 61:2. 265-266.
Kostoff, R. N. (2003d). The practice and malpractice of stemming. JASIST. 54:10. 984-985. August.
Kostoff, R.N., Shlesinger, M.F., Malpohl, G. (2004a). Fractals roadmaps using bibliometrics and database tomography. Fractals. 12:1. 1-16.

Kostoff, R.N., Shlesinger, M., and Tshiteya, R. (2004b). Nonlinear dynamics roadmaps using bibliometrics and Database Tomography. International Journal of Bifurcation and Chaos. 14:1. 61-92.
Kostoff, R.N., Bedford, C.W., Del Rio, J. A., Cortes, H., and Karypis, G. (2004c). Macromolecule mass spectrometry: Citation mining of user documents. Journal of the American Society for Mass Spectrometry. 15:3. 281-287. March.
Kostoff, R.N., and Shlesinger, M.F. (2004d). CAB-Citation-assisted background. Scientometrics. 62:2. 199-212. 2005.

Kuhnhold, M. (2000). The concept of "text mining". Wirtschaftsinformatik, 42 (2). 175-179. Apr.
Lagus, K., Honkela, T., Kaski, S., and Kohonen, T. (1999). Websom for textual data mining. Artificial Intelligence Review. 13(5-6). 345-364. December.
Losiewicz, P., Oard, D., and Kostoff, R.N. (2000).Textual Data Mining to Support Science and Technology Management. Journal of Intelligent Information Systems. 15.
McArdle, J.J. Principles versus Principals of Structural Factor-Analyses. Multivariate Behavioral Research, 25 (1). 81-87. Jan 1990.
Narin, F. (1976). Evaluative bibliometrics: the use of publication and citation analysis in the evaluation of scientific activity (monograph). NSF C-637. National Science Foundation. Contract NSF C-627. NTIS Accession No. PB252339/AS.
Narin, F., Olivastro, D., Stevens, K.A. (1994). Bibliometrics theory, practice and problems. Evaluation Review. 18(1). 65-76.
Perrin, P. and Petry, F.E. (2003). Extraction and representation of contextual information for knowledge discovery in texts. Information Sciences, 151. 125-152. May.
Porter, M.F. (1980). An algorithm for suffix stripping. Program, 14(3). 130-137.
Rasmussen, E. (1992). *Clustering Algorithms*. In W. B. Frakes and R. Baeza-Yates (eds.). Information Retrieval Data Structures and Algorithms, Prentice Hall, N. J.
Salton, G. and Buckley, C. (1998). Term-weighting approaches in automatic text retrieval. Information Processing and Management. 24:5. 513-523.
Schenker, A., Last, M., Bunke, H., and Kandel, A. (2003). Graph representations for web document clustering. Pattern Recognition and Image Analysis, Proceedings, 2652. 935-942.

Schroeder, M. (1991). Fractals, chaos, power laws: Minutes from an infinite paradise. (W.H. Freeman, New York, NY).
Schubert, A., Glanzel, W., Braun, T. (1987). Subject field characteristic citation scores and scales for assessing research performance. Scientometrics. 12 (5-6): 267-291.
Stegmann, J., and Grohmann, G. (2003). Hypothesis generation guided by co-word clustering. Scientometrics, 56 (1). 111-135.
Steinbach, M., Karypis, G., and Kumar, V. A. (2000). Comparison of document clustering techniques. Technical Report #00--034. Department of Computer Science and Engineering. University of Minnesota.
Stensmo, M. (2002). A scalable and efficient probabilistic information retrieval and text mining system. Artificial Neural Networks. ICANN 2002. 2415. 643-648.
Swanson D.R. (1986). Fish oil, Raynaud's Syndrome, and undiscovered public knowledge. Perspectives in Biology and Medicine. 30 (1). 7-18. Fall.
Swanson, D.R., Smalheiser, N.R. (1997). An interactive system for finding complementary literatures: a stimulus to scientific discovery. Artificial Intelligence. 91 (2). 183-203.
TREC (2003). (Text Retrieval Conference), Home Page, http://trec.nist.gov/.
Trybula, W.J. (1999). Text mining. Annual Review of Information Science and Technology. 34. 385-419.
Viator, J.A., Pestorius, F.M . (2001). Investigating trends in acoustics research from 1970-1999. Journal of the Acoustical Society of America. 109 (5): 1779-1783 Part 1.
Visa, A. (2001). Technology of text mining. Machine Learning and Data Mining in Pattern Recognition. 2123. 1-11.
Wang, B.B., McKay, R.I., Abbass, H.A., and Barlow, M. (2003). A comparative study for domain ontology guided feature extraction. Twenty-Fifth Australian Computer Science Conference [ACSC2003], Adelaide, Australia, Vol. 16, Michael Oudshoorn, ed.
Weeber M., Klein H., Aronson A.R,, Mork J.G, de Jong-van den Berg, L.T.W, and Vos R. (2000). Text-based discovery in biomedicine: The architecture of the DAD-system. Journal of the American Medical Informatics Association. 903-907 Suppl. S.
Weiss, S.M., Apte, C., Damerau, F.J., Johnson, D.E., Oles, F.J., Goetz, T., and Hampp, T. (1999). Maximizing text-mining performance. IEEE Intelligent Systems & Their Applications. 14:4. 63-69. July-August.

Wilbur W. J. and Sirotkin K. (1992). The automatic identification of stop words. Journal of Information Science. 18 (1). 45-55.
Wilbur W.J. and Yang Y.M. (1996). An analysis of statistical term strength and its use in the indexing and retrieval of molecular biology texts. Computers in Biology and Medicine. 26 (3). 209-222. May.
Willet, P. (1988). Recent trends in hierarchical document clustering: A critical review. *Information Processing and Management.* 24:577-597.
Yalcin, I. and Amemiya, Y. (2001). Nonlinear factor analysis as a statistical method. Statistical Science, 16 (3). 275-294. Aug.
Zamir, O. and Etzioni, O. (1998). Web document clustering: A feasibility demonstration. In: *Proceedings of the 19th International ACM SIGIR Conference on Research and Development in Information Retrieval* (SIGIR'98). 46-54.
Zhu, D.H. and Porter, A.L. (2002). Automated extraction and visualization of information for technological intelligence and forecasting. Technological Forecasting and Social Change. 69:5. 495-506. June.

APPENDIX 1 – SCREE PLOT INTERPRETATION

A parametric analysis of factor number determination as a function of Scree Plot scale was performed, to test whether factor number determination was scale dependent. The results follow.

1) Demonstration of Fractality

Factor matrices with different numbers of factors specified were computed. Eigenvalues were generated by Principal Components Analysis, and these eigenvalues represented the variance accounted for by each underlying factor. Figure 1 shows the factor eigenvalue-factor number plot for the 659 un-rotated factors on a linear scale. The 'elbow', or break point, of the curve appears to be about fourteen factors. To improve resolution, the curve was stretched in the x direction by halving the number of factors shown on one page. The curve had a similar shape to the 659 factor case, but the factor termination point appeared to decrease. The halving process was repeated until ten factors were plotted on one page, and the resolution effectively increased by an order of magnitude overall.

Figure 3 shows the ten factor plot. The elbow of the curve appears to be about two factors. Thus, the number of factors selected based on significant

slope change decreased from fourteen in the 659 factor plot to two in the ten factor plot.

INSERT FIGURE 3

In fractal analysis, a fractal object has a number of characteristics (Schroeder, 1991). Among these are self-similarity (similar to itself at different magnifications), and adherence to a scaling relationship (the measured value of a property will depend on the resolution used to make the measurement). The Scree Plot appears to have these two fractal properties. As the resolution increases, more structure appears, and the value of the break point changes.

The simplest and most common form of the scaling relationship is that of a power law. When such a power law is plotted on a log-log scale, the scaling relationship appears as a straight line. Figure 4 is a plot of the break point on a linear scale, and Figure 5 is a re-plot of Figure 4 on a log-log scale. The log-log plot is approximately linear, reflects power law scaling, and validates the break point selection as a fractal process. This observation about the fractal-like nature of the Scree Plot analysis process does not appear to have been reported in the literature previously. Further, this sensitivity of factor number to scale confirms the reviewer's concerns about the precision of the Scree Plot approach to determining number of factors to be selected.

INSERT FIGURE 4

INSERT FIGURE 5

The reviewer also raised the concern about the impact of the inconsistency among analysts in interpreting the Scree Plot on the final taxonomy. Essentially, this issue concerns the consequences of selecting a few more or less factors to run.

The considerations involved in selecting the number of factors to run are similar to those involved in selecting the number of clusters for the cluster analysis. Basically, the issue revolves around the level of resolution desired. More clusters, or more factors, provide more resolution, detail, and information. In addition, more factors capture a greater fraction of total variance. The cost of more factors or clusters is more computer running time, and especially more time for interpretation of results. Thus, the number of clusters and/ or factors to be selected depends on the objectives of the study.

As an example of the need for resolution, the first author has performed text mining of both homogeneous databases and heterogeneous databases in the past few years. The homogeneous databases derive from monodiscipline studies, where the topical material is closely related (e.g., the anthrax database). The heterogeneous databases derive from multi-discipline databases, where the topical material can be very disparate (e.g., China's research output).

The single discipline studies require relatively few clusters to capture the main themes of the database, because the topical variation is modest, and relatively few hierarchical taxonomy levels are required for the same reason. The multi-discipline studies require many clusters to resolve the disparate themes, and many taxonomy levels may be required to portray the structure accurately.

For example, suppose it were desired to examine the research outputs of a major country whose research budget was one billion dollars per year, whose research program consisted of 100 different disciplines, and whose research output database consisted of 10000 records. If two clusters were used for the analysis, only one hierarchical taxonomy level would be possible, and each cluster (on average) would cover 500 million dollars per year worth of research, fifty technical disciplines, and 5000 records. Obviously, the results and corresponding insights would be very generic and aggregated, and probably of very limited utility. If, however, 1000 clusters were used, then ten hierarchical taxonomy levels would be possible, and each elemental cluster (on average) would cover one million dollars worth of research, 1/10 of a technical discipline, and ten records. Much more detailed understanding of the country's research would be obtained, down to approximately the individual program level. Obviously, much more work

would be required to generate useful information from the 1000 cluster case than the two cluster case.

In the present study, the issue of factor number variations was examined by the authors in the initial phase of the study by performing a parametric study of number of factors verses taxonomy structure. Factor matrices ranging from two factors in size to fourteen factors were run, and the results analyzed. The main themes did not change, and the context-dependent trivial words identified remained essentially the same. The main changes were the number of hierarchical levels that could be generated, and the resolution afforded by each factor.

For the two factor and fourteen factor taxonomies, the main themes are still autoimmunity and circulation. The fourteen factor taxonomy allows more structural detail to be shown, as displayed in the Results section. More levels in the taxonomy could have been generated with the fourteen factor taxonomy, if desired, and more sub-themes could have been generated with more detailed specificity. Obviously, if there had been three main themes instead of two for this topical area, then the two factor case would have missed one of the themes. In practice, the analyst should always perform some type of sensitivity study on number of factors, to insure that no main themes are being missed with the final number of factors chosen.

APPENDIX 2 – FACTOR MATRIX FILTERING

The purpose of this Appendix is to show how two word clusters, generated using the WINSTAT multi-link hierarchical aggregation technique, can be compared. The comparison method is a combination of quantitative and qualitative approaches.

This Appendix contains two sections. The first section shows the impact on the ‘distance’ metric, and on the assignment of words to clusters, of substituting the most trivial words for higher technical content words in the selection of cluster input words. The second section shows the impact on the ‘distance’ metric, and on the assignment of words to clusters, of substituting non-factor matrix-selected words by factor matrix-selected words in the selection of cluster input words.

Impact of Trivial Words on Cluster Quality

The impact was shown by conducting a simple experiment where trivial words were substituted for higher technical content words, and the change in cluster quality evaluated. The same 930 record database that was used in the main text of the study was used in this Appendix, for consistency. The words contained in the Abstract were extracted using the TechOasis software package, subject to the standard StopWord list filtering. This StopWord list contains words that are trivial in almost every context (e.g., the, if, or, and, etc).

In the first case, the 252 highest frequency words extracted from the Abstract were used as input for the WINSTAT clustering process. In the second case, the two highest frequency trivial words on the StopWord list (of, the) were substituted for the two lowest frequency words of the 252 words used for the first case (abnormalities, C). While these two words are not of the highest technical content, they are certainly of higher content than 'of', 'the'.

Clusters were run using the 8, 16, 32, 64, 128, and 253 highest frequency words, respectively. For display purposes, the 64 word dendrograms were selected. Figure 6 shows the non-trivial word dendrogram, while Figure 7 shows the dendrogram that includes the two trivial words. The overall 'distance' metric value associated with Figure 6 is 132.56, while the overall 'distance' metric value associated with Figure 7 is 139.54. For the 252 word cases, the respective overall 'distance' metric values are 513.81 and 596.32. As expected, substitution of the trivial words 'of', 'the' for the higher technical content words 'abnormalities', 'C' results in an increase in the value of the 'distance' metric. As the constituent clusters become more diffuse with the addition of trivial words, the 'distance', a measure of cluster cohesiveness, increases.

INSERT FIGURE 6

INSERT FIGURE 7

Of equal importance is the impact on assignment of words to clusters. The cluster thematic differences are most pronounced at the highest taxonomy levels, and these differences become more subtle as the lowest taxonomy levels are accessed. As shown in the text, and known from medical experience, the first taxonomy level of the Raynaud's Phenomenon literature has the two major categories of auto-immunity and circulation. The

literatures, and key phrases, associated with each category tend to be distinct. How did the substitution of trivial words for higher content words impact the assignment of words to the highest level categories in the taxonomy?

On Figure 6, there are two obvious first level clusters. One ranges from the words Raynaud's to Diagnosis, and the other ranges from the words Blood to Disorders. The Raynaud's-Diagnosis cluster focuses on Auto-immunity, while the Blood-Diagnosis cluster focuses on Circulation. The words in the clusters range from very specific (e.g., lupus, scleroderma, vascular, arterial) to quite general (e.g., test, years, cases, severe). The presence of the general words is a consequence of not using any filtering (manual, factor matrix, etc) for the demonstration in this section, other than the StopWord list and the frequency cut-off.

On Figure 6, the first level Auto-immunity category contains 31 words (48%), and the first level Circulation category contains 33 words (52%). The key auto-immunity single words and obvious word combinations (e.g., systemic sclerosis, lupus erythematosus) are in the Auto-immunity category, and the key circulation single words and obvious word combinations (e.g., blood flow, finger temperature) are in the Circulation category.

On Figure 7, there is a major change in the taxonomy structure. The first level of the taxonomy splits into two categories. One is a very small category, ranging from Of to Clinical (in abbreviated form, Of-Clinical), and the other is a much larger category (Systemic-Disorders). The small category is a very generic overview of the database. The larger category combines the Auto-immunity and Circulation categories.

The larger category splits into two second level sub-categories. One has an Auto-immunity focus (Systemic-3), and the other has a Circulation focus (Blood-Disorders). The Auto-immunity sub-category contains 33 words (59%), while the Circulation sub-category contains 23 words (41%). Thus, for the major technical sub-division between Auto-immunity and Circulation, Auto-immunity has gone from 48% of the words to 59% when the trivial words are substituted, while Circulation has decreased from 52% to 41%, a noticeable change.

What are the words that switched categories? Most were generic, concentrated in a small Circulation cluster near the far end of the

dendrogram on Figure 6 (Group-3). The other two were more specific (peripheral, vascular). They switched from the Circulation category on Figure 6 to the Auto-immunity category on Figure 7. While it could be argued that the generic terms that switched categories were only weakly linked thematically to the Circulation category initially, the switch of the more specific terms (peripheral, vascular) is less defendable.

There are 52 records in the 930 Raynaud's record database that contain the words peripheral and vascular. A sampling of the 52 records shows that the combination of peripheral and vascular is always used in the context of circulation. For primary Raynaud's, there is only the association with circulation. For secondary Raynaud's, where circulatory problems may be a symptomatic spin-off of the auto-immune disease, such articles might include the underlying auto-immune disease in conjunction with the circulatory problem. However, the main association of the peripheral-vascular combination is with circulation problems. Thus, in addition to changing the first level structure of the taxonomy, an additional thematic effect of substituting the two trivial words into the clustering input has been to re-assign the combination peripheral and vascular from the more appropriate technical category to the less appropriate category.

The number of trivial words used in this substitution experiment was small (two). As the number of words used to form the clusters increases, the effect of the two substituted trivial words on the results is expected to decrease. In the 252 word case that includes the substituted trivial words, the combination peripheral and vascular reverts to the Circulation category in the clusters with the substituted trivial words. However, the separate generic first level category containing the substituted trivial words and the more generic words remains.

Additionally, in the 252 word case that includes the substituted trivial words, the combination vascular and peripheral reverts to the appropriate high level Circulation category. However, the individual words vascular and peripheral are in different lower level categories relative to the 252 word case that does not include the substituted trivial words. In the latter case, peripheral and vascular are adjacent. Thus, the simple substitution of two trivial words has changed the taxonomy structure at the highest level, and more pervasively at the lowest levels.

Impact of Factor Matrix-Selected Words on Cluster Quality

The impact was shown by comparing the word dendrogram from the main text (whose input words were obtained with factor matrix filtering) with a word dendrogram whose input words were not obtained with factor matrix filtering. The comparison results will depend strongly on the method used for word selection in the latter case. To minimize human intervention, and the potential for arbitrary bias, the highest frequency words after StopWord list filtering were selected for clustering, including conflation of closely related words.

The 252 word dendrograms are shown on Figures 2 and 8. Figure 2 incorporates factor matrix filtering, while Figure 8 does not. The overall distance metric associated with Figure 2 is 513.06, while the overall distance metric associated with Figure 8 is 527.28. For the 64 word cases, the respective 'distance' metric values are 134.08 and 140.5. As expected, substitution of factor matrix filtering for non-factor matrix filtering results in a decrease in the value of the 'distance' metric. As the constituent clusters become sharper with the removal of context-dependent trivial words from factor matrix filtering, the 'distance', a measure of cluster cohesiveness, decreases.

INSERT FIGURE 8

As in the previous example, the assignment of words to clusters is of equal importance to the change is particular metrics. How did the addition of factor matrix filtering impact the assignment of words to the highest level categories in the taxonomy?

To re-iterate the structure of Figure 2, there are the two first level categories of Auto-immunity (Patients-Old) and Circulation (Treatment-Acid). Auto-immunity can be subdivided into its second level categories of Auto-immune Diseases/ Antibodies (Patients-Neuropathy) and Inflammation (Lesions-Old), and Circulation can be sub-divided into its second level categories of peripheral vascular circulation (Treatment-Consecutive) and coronary vascular circulation (Pulmonary-Acid). The categories are sharp, correspond to medical diagnosis and treatment, and the contents are appropriately placed.

Figure 8 contains a different higher category level structure. The first level can be divided into two categories, a very generic small category (13 words) centered around Raynaud's Phenomenon and scleroderma (Raynaud's-Symptoms), and the other being large (239 words) and including both Auto-immunity and Circulation (Systemic-Abnormality). The small generic first level category can be subdivided into its second level categories centered around Raynaud's Phenomenon/ Scleroderma (Raynaud's-Severity) and very generic terms (One-Symptoms). The larger second level category can be sub-divided into its second level categories of Auto-immunity (Systemic-Signs) and Circulation (Treatment-Abnormality).

Not only is the structure of the first level different between the two figures, but, for example, the word 'scleroderma' is in different first level categories. In the factor matrix filtered case (Figure 2), scleroderma is in the Auto-immunity category, closely linked to both relatively generic terms (patients, diseases) and a very specific term (progressive systemic sclerosis, PSS). This is due to 1) PSS being a pseudonym for scleroderma and 2) scleroderma being a sign of a group of diseases that involve the abnormal growth of connective tissue, which supports the skin and internal organs, and therefore used as a more generic umbrella term for these disorders. In the non-factor matrix filtered case (Figure 8), scleroderma is in the generic first level category, coupled to the generic terms 'patients, diseases', as in Figure 2, but de-coupled from its pseudonym PSS.

A third level sub-division is required for Figure 8 for comparison with the second level sub-division of Figure 2. The second level Auto-immunity category of Figure 8 (Systemic-Signs) sub-divides into third level categories of Auto-immune Diseases/ Antibodies (Systemic-Comparison) and Inflammation (Lesions-Signs). The Inflammation category contains 13 words, of which perhaps 3 relate specifically to inflammation (lesions, vasculitis, inflammatory). Contrast this with the second level Inflammation category from Figure 2. This Inflammation category contains 16 words (similar in magnitude), of which 8 relate specifically to inflammation (lesions, corticosteroids, eosinophilia, fasciitis, hypergammaglobulinemia, scleroderma-like, inflammatory, polyarthritis). The difference in level of detail is striking.

The second level Circulation category of Figure 8 (Treatment-Abnormality) sub-divided into third level categories of Circulation (Treatment-Data) and a very generic unfocused category (Criteria-Abnormality). In the third level

Circulation category, peripheral vascular circulation is intermingled with coronary artery circulation, and lower taxonomy levels need to be accessed before these circulation sub-categories can be differentiated.

On Figure 8, the circulation sub-category of Coronary Artery Circulation (Combined-Five) contains 12 terms, of which 3 are relatively specific (hypertensive, cardiac, heart). On Figure 2, the second level category of Coronary Artery Circulation (Pulmonary-Acid) contains 25 terms, of which about 20 are relatively specific (pulmonary, fibrosis, hypertension, cardiac, cardiovascular, heart, coronary, myocardial, ventricular, angina, necrosis, spasm, chest, lung, biofeedback, training, relaxation, stress, migraine, headaches). The biofeedback thrust was not even mentioned on Figure 8. Additionally, Figure 2 provides much more detail about the physical consequences of insufficient coronary artery circulation, whereas Figure 8 alludes to the general area with no specific detail.

There are many other differences in structure and content between the two dendrograms. In summary, the factor matrix filtering provides a lower value of overall ‘distance’ (translating into a more sharply defined overall taxonomy. The clusters are improved from a medical perspective, and the contents are far more detailed.

Figure 1

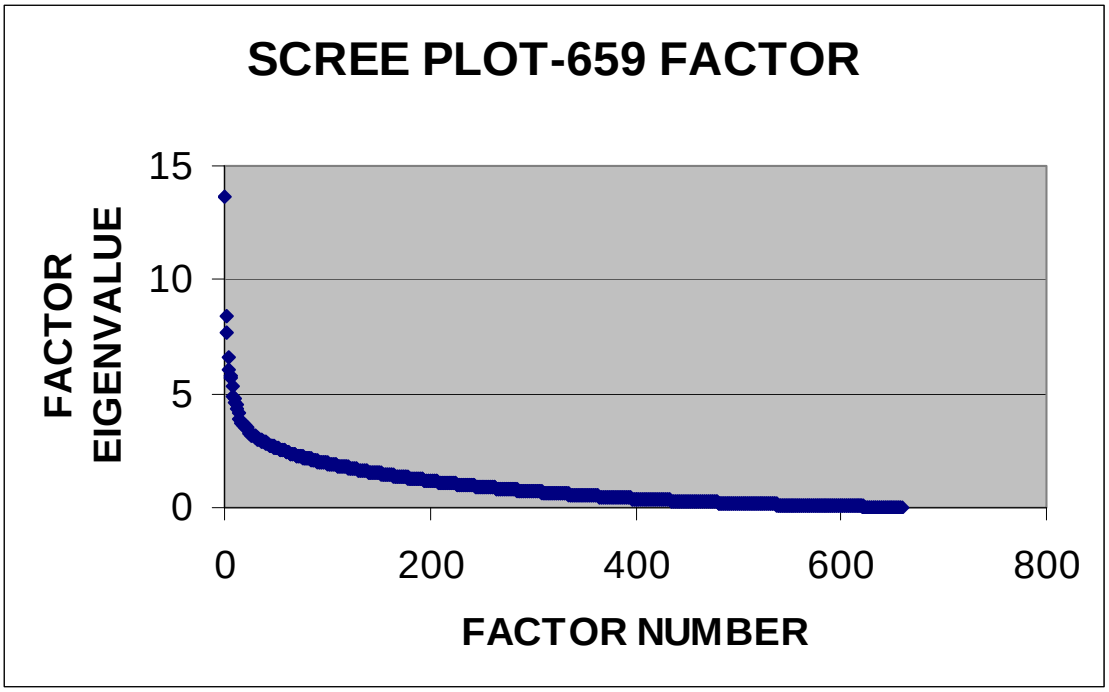

SCREE PLOT-659 FACTOR
FACTOR EIGENVALUE
15
10
5
0
0
200
400
600
800
FACTOR NUMBER

Figure 2

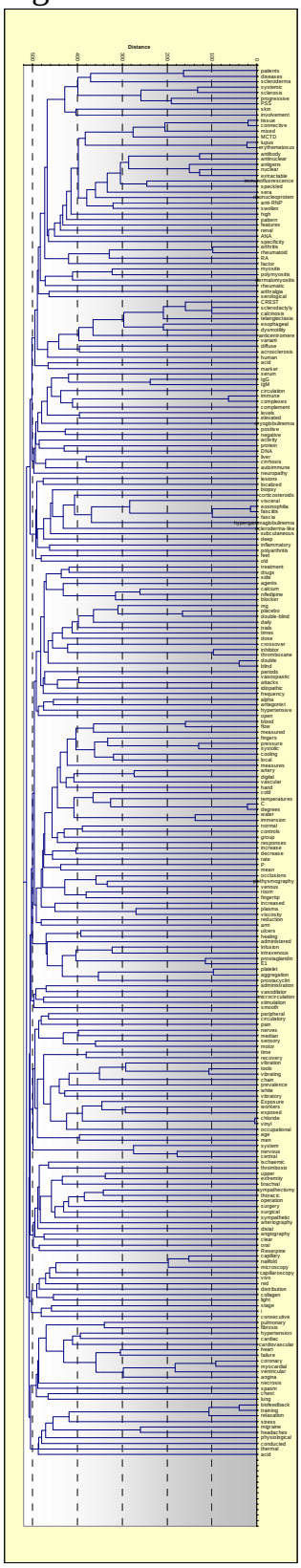

Distance
500
400
300
200
100
0
patients
disease
scleroderma
systemic
sclerosis
progressive
skin
virus
involvement
tissue
connective
mixed
cases
lupus
erythematosus
antibody
antinuclear
nuclear
antigen
extractable
immunofluorescence
speckled
sera
chromosprotein
centromere
positive
high
pattern
features
titers
ANA
specificity
arthritis
rheumatoid
RA
tissue
myositis
polymyositis
dermatomyositis
rheumatic
arthralgia
overlap
CREST
sclerodactyly
calcinosis
telangiectasia
Raynaud
dysmotility
anticentromere
variant
diffuse
scleroderma
human
sera
marker
serum
IgG
CPK
circulation
immune
complexes
complement
levels
elevated
hypergammaglobulinemia
positive
negative
activity
protein
DNA
lung
cirrhosis
autoimmune
neuropathy
lesions
elevated
biopsy
corticosteroids
visceral
eosinophilia
fascia
fasciitis
hypergammaglobulinemia
scleroderma-like
subcutaneous
deep
inflammatory
polyarthritis
liver
old
treatment
drugs
blood
agents
calcium
nifedipine
blocker
CA
placebo
double-blind
daily
trials
doses
oral
crossover
inhibitor
thromboxane
double
blind
placebo
eosinophilic
attacks
idiopathic
frequency
attack
antagonist
hypertensive
open
blood
flow
measured
mean
pressure
systolic
cooling
local
responses
artery
digital
vascular
hand
cold
temperatures
C
degrees
water
immersion
normal
controls
group
responses
increase
decrease
site
C
mean
occlusions
plethysmography
venous
C
fingertip
increased
plasma
viscosity
reduction
C
ulcers
healing
administered
infusion
intravenous
prostaglandin
E1
C
platelet
aggregation
inhibition
administration
vasodilator
microcirculation
stimulation
smooth
peripheral
circulatory
pain
nerves
median
sensory
C
time
recovery
vibration
tools
vibrating
prevalence
white
vibratory
Exposure
workers
exposed
chloride
vinyl
occupational
age
C
system
nervous
central
ischaemic
thrombosis
upper
extremity
brachial
sympathectomy
thoracic
operation
C
surgical
sympathetic
arteriography
distal
digital
C
chair
oral
fluorescein
capillary
nailbed
microscopy
microscopy
vessels
red
distribution
collagen
C
stage
I
consecutive
pulmonary
C
hypertension
cardiac
cardiovascular
heart
failure
coronary
myocardial
ventricular
angina
necrosis
spasm
chest
lung
biofeedback
training
relaxation
C
migraine
headaches
physiological
conducted
thermal
acid

Figure 3

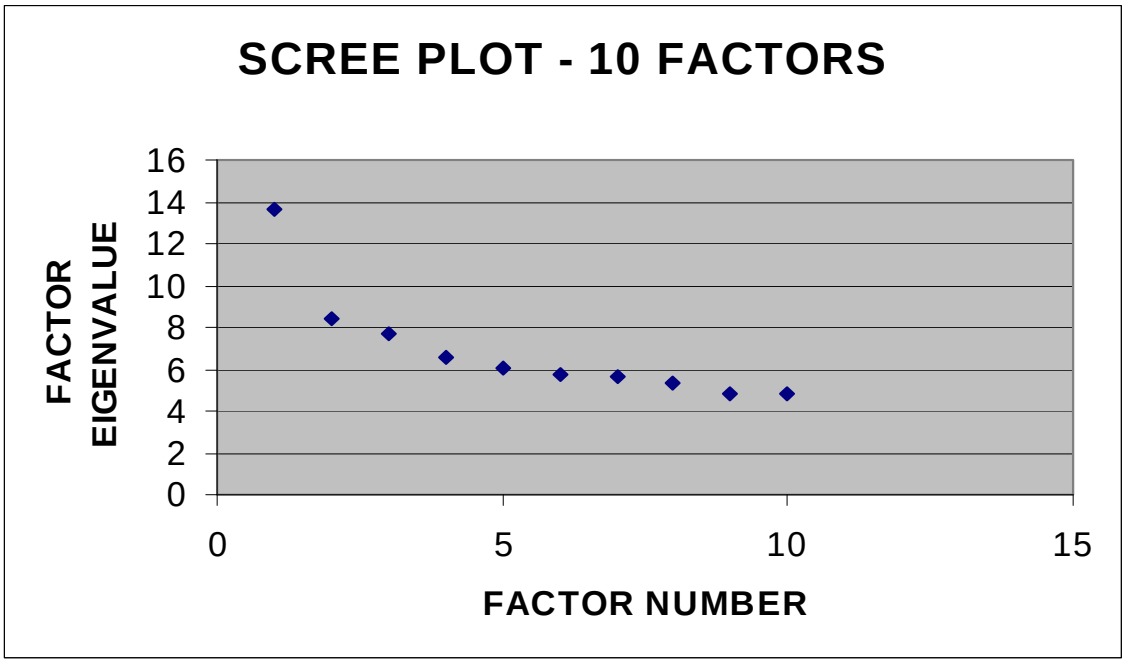

SCREE PLOT - 10 FACTORS
FACTOR EIGENVALUE
16
14
12
10
8
6
4
2
0
0
5
10
15
FACTOR NUMBER

Figure 4

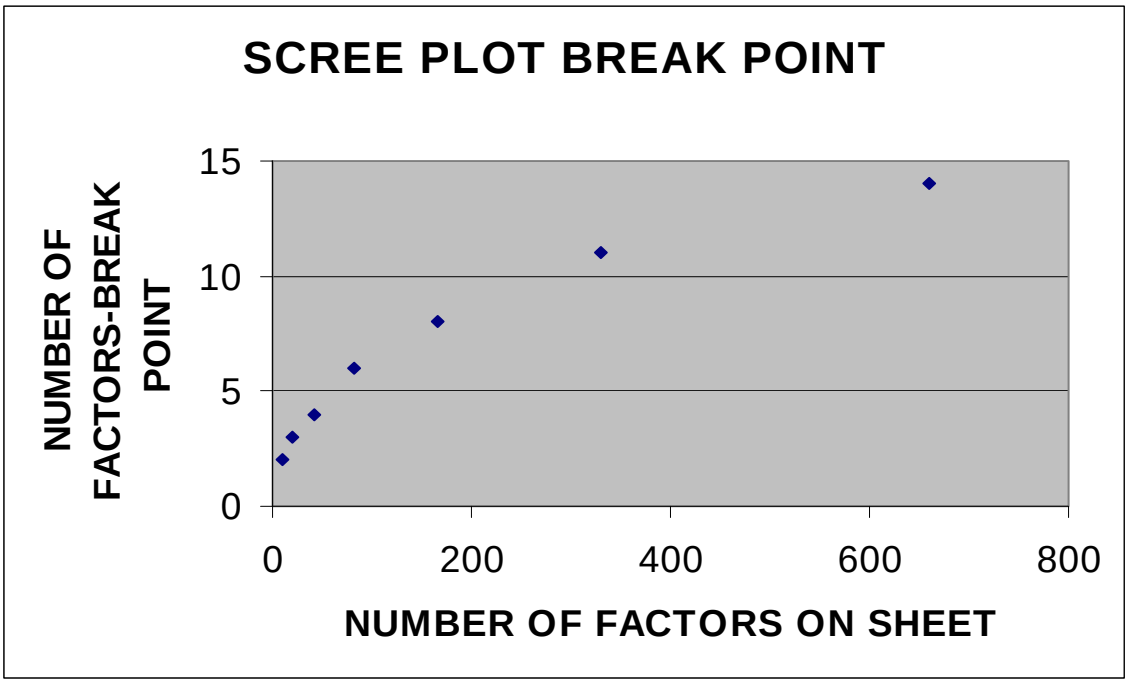

SCREE PLOT BREAK POINT
NUMBER OF FACTORS-BREAK POINT
15
10
5
0
0
200
400
600
800
NUMBER OF FACTORS ON SHEET

Figure 5

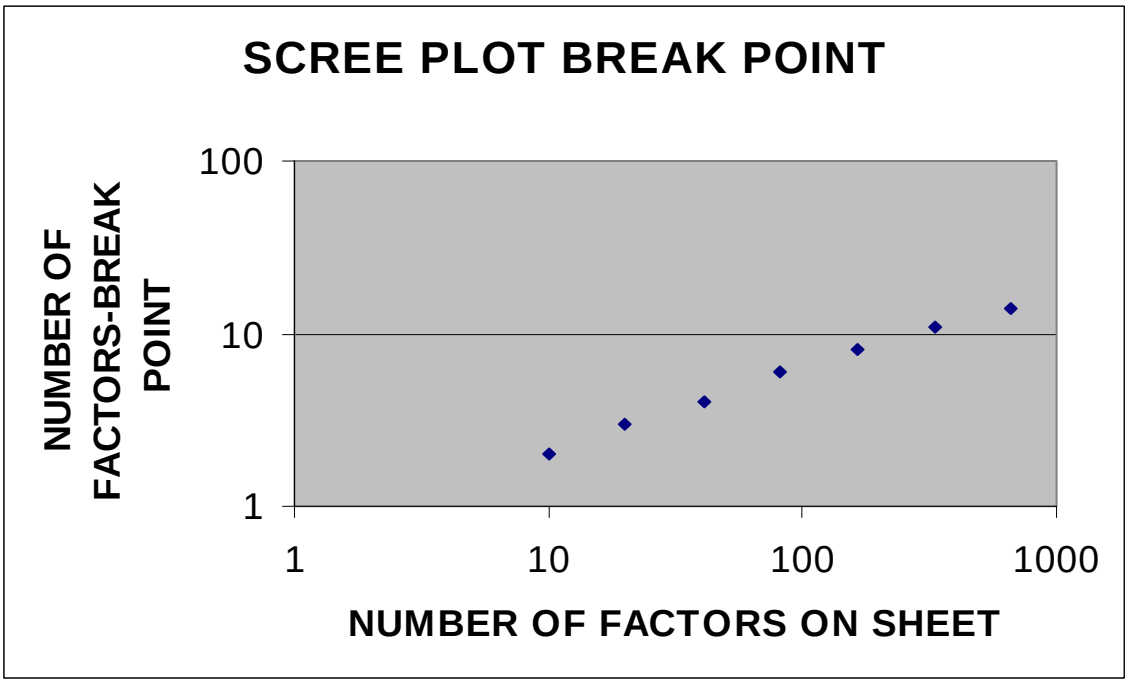

SCREE PLOT BREAK POINT
NUMBER OF FACTORS-BREAK POINT
100
10
1
1
10
100
1000
NUMBER OF FACTORS ON SHEET

Figure 6

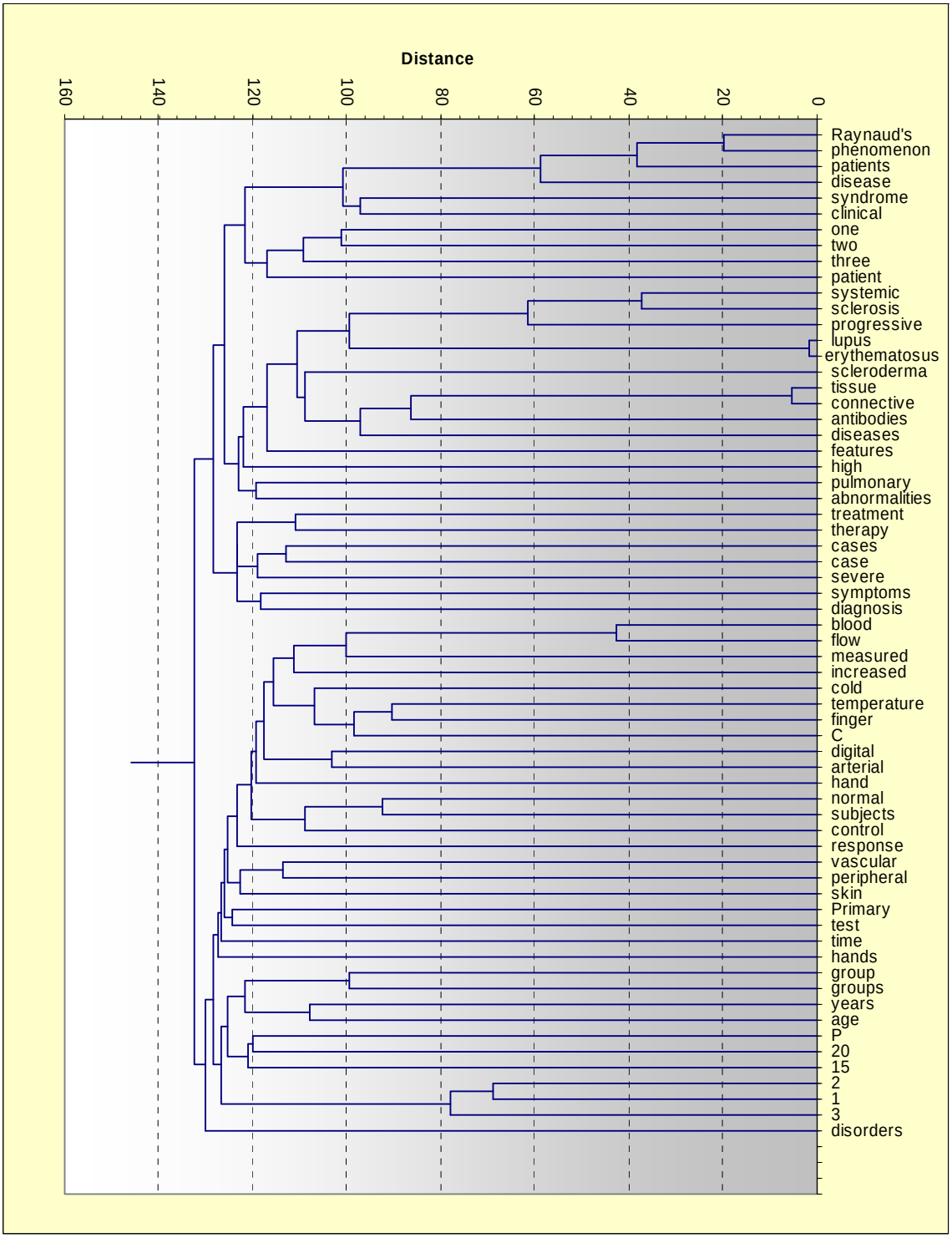

Distance
160
140
120
100
80
60
40
20
0
Raynaud's
phenomenon
patients
disease
syndrome
clinical
one
two
three
patient
systemic
sclerosis
progressive
lupus
erythematosus
scleroderma
tissue
connective
antibodies
diseases
features
high
pulmonary
abnormalities
treatment
therapy
cases
case
severe
symptoms
diagnosis
blood
flow
measured
increased
cold
temperature
finger
C
digital
arterial
hand
normal
subjects
control
response
vascular
peripheral
skin
Primary
test
time
hands
group
groups
years
age
P
20
15
2
1
3
disorders

Figure 7

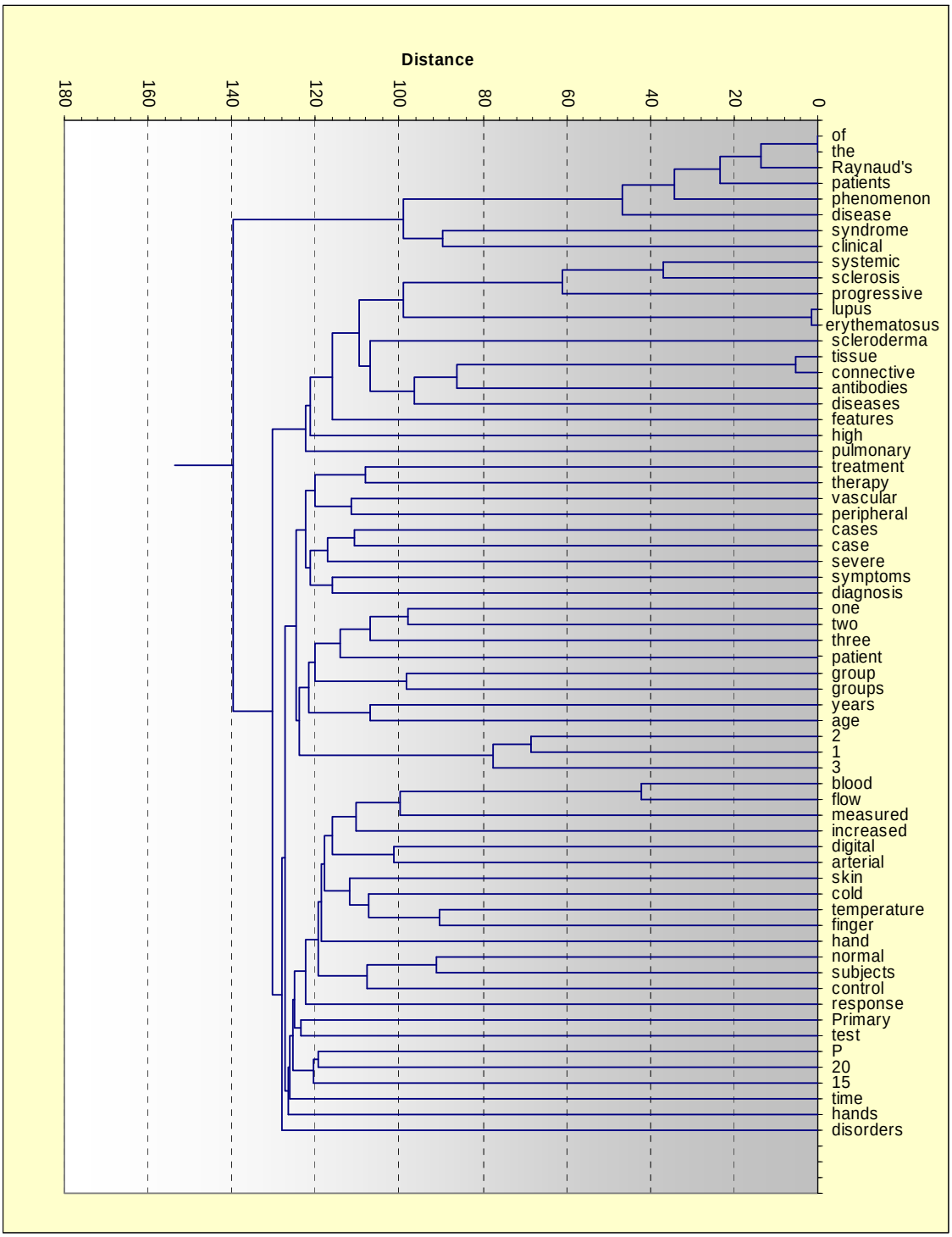
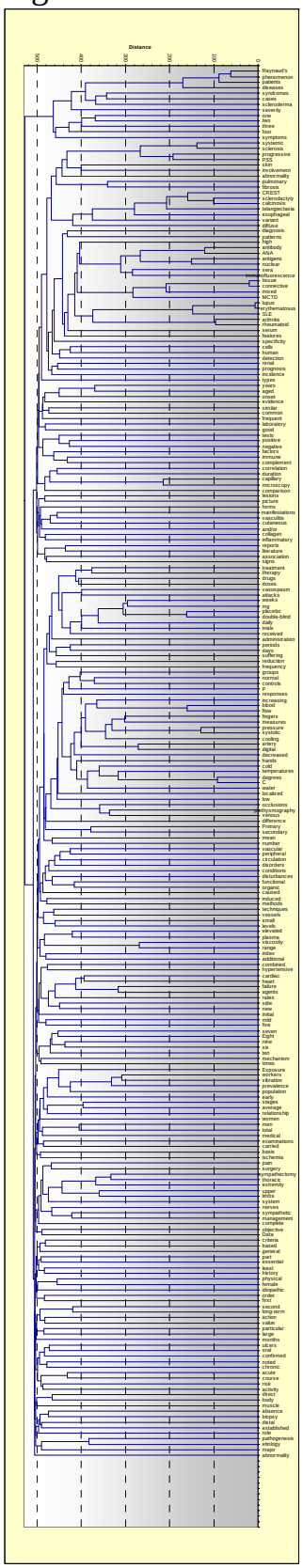

Distance
180
160
140
120
100
80
60
40
20
0
of
the
Raynaud's
patients
phenomenon
disease
syndrome
clinical
systemic
sclerosis
progressive
lupus
erythematosus
scleroderma
tissue
connective
antibodies
diseases
features
high
pulmonary
treatment
therapy
vascular
peripheral
cases
case
severe
symptoms
diagnosis
one
two
three
patient
group
groups
years
age
2
1
3
blood
flow
measured
increased
digital
arterial
skin
cold
temperature
finger
hand
normal
subjects
control
response
Primary
test
P
20
15
time
hands
disorders

Figure 8


Distance
500
400
300
200
100
0
Raynaud's
phenomenon
disorders
diseases
syndromes
cases
scleroderma
severity
one
two
three
four
symptoms
systemic
sclerosis
progressive
PSS
skin
involvement
abnormality
pulmonary
CREST
sclerodactyly
calcinosis
telangiectasia
esophageal
variant
SSc
diagnosis
patterns
high
antibody
ANA
antigens
nuclear
area
fluorescence
tissue
connective
CTD
MCTD
lupus
erythematosus
SLE
profile
rheumatoid
serum
features
specificity
cells
human
infection
renal
prognosis
incidence
types
years
aged
male
evidence
similar
common
frequent
laboratory
good
tests
positive
negative
factors
disease
complement
correlation
duration
capillary
microscopy
comparison
incidence
picture
forms
manifestations
vasculitis
cutaneous
ulcers
collagen
inflammatory
reports
literature
association
signs
cause
treatment
therapy
drugs
doses
vasospasm
attacks
weeks
mg
placebo
double-blind
daily
trial
received
administration
periods
days
suffering
decreased
frequency
groups
normal
controls
P
responses
following
blood
flow
fingers
measures
pressure
systolic
cooling
artery
digital
decreased
hands
cold
temperatures
degrees
C
water
localized
skin
occlusions
plethysmography
venous
difference
Primary
Secondary
mean
number
vascular
peripheral
circulation
conditions
disorders
disturbances
functional
organic
caused
clinical
methods
techniques
vessels
small
levels
elevated
plasma
viscosity
range
index
additional
combined
hypertensive
cardiac
heart
failure
agents
rates
rate
male
initial
mild
five
seven
eight
four
six
ten
mechanism
times
exposure
workers
vibration
prevalence
population
early
stages
change
relationship
women
men
total
medical
examinations
control
basis
ischemia
pain
surgery
sympathectomy
thoracic
extremity
upper
limbs
system
nervous
sympathetic
management
complete
objective
Data
study
based
general
part
essential
least
theory
critical
female
idiopathic
order
first
proposed
long-term
action
value
particular
large
months
small
confirmed
noted
chronic
acute
course
may
activity
direct
body
muscle
tissue
biopsy
distal
established
role
pathogenesis
etiology
major
abnormality

Section 8. Use and Misuse of Citation Analysis in Text Mining.

(based on Kostoff, R. N. "The Use and Misuse of Citation Analysis in Research Evaluation". Scientometrics. 43:1. September 1998.)

The present section addresses some of the many possible uses of citations, including bookmark, intellectual heritage, impact tracker, and self-serving purposes. The main focus is on the applicability of citation analysis as an impact or quality measure. If a paper's bibliography is viewed as consisting of a directed (research impact or quality) component related to intellectual heritage and random components related to specific self-interest topics, then for large numbers of citations from many different citing papers, the most significant intellectual heritage (research impact or quality) citations will aggregate and the random author-specific self-serving citations will be scattered and not accumulate. However, there are at least two limitations to this model of citation analysis for stand-alone use as a measure of research impact or quality. First, the reference to intellectual heritage could be positive or negative. Second , there could be systemic biases which affect the aggregate results, and one of these, the "Pied Piper Effect", is described in detail. Finally, the results of a short citation study comparing Russian and American papers in different technical fields are presented. The questions raised in interpreting this data highlight a few of the difficulties in attempting to interpret citation results without supplementary information.

Leydesdorff (Leydesdorff, 1998) addresses the history of citations and citation analysis, and the transformation of a reference mechanism into a purportedly quantitative measure of research impact/ quality. The present paper examines different facets of citations and citation analysis, and discusses the validity of citation analysis as a useful measure of research impact/ quality.

1. Citations

1.1. Citations as bookmarks

The starting point for this section centers around the need for citations. Why do we use citations in a paper? There are obviously many reasons for citations, ranging from contributions to the advancement of science and knowledge to less noble purposes for inclusion in text. Some of these reasons will be enumerated in the following paragraphs.

We start with the bookmark function of citations. The average reader of a technical paper typically does not have the luxury to expend large amounts of time on extracting useful information from the paper. The shorter the paper, the greater is the likelihood that it will be read in its entirety. Citations, like acronyms or mathematical symbols or 'laws', provide a condensed reference to a much larger body of data. The relatively few readers who would be interested in such details can examine them at a later date.

I could write a paper including Lotka's law without providing a reference to Lotka's law, or without even mentioning the name 'Lotka's law'. Whenever the need to include Lotka's law arose, I would write out the definition. This unabridged approach to writing would lead to an unnecessarily lengthy document, and would lose the average reader quite rapidly. Using the abridged description 'Lotka's law' allows for an efficiency of presentation. Including such a citation allows the reader to access more details, shows evidence of my awareness of other related works, and probably provides more credibility to my paper in the reader's eyes.

1.2. Citations as intellectual heritage linkages

Other than the shorthand function, citations provide links to the intellectual heritage foundation for the citing paper, and help provide the historical context for displaying the unique contributions of the citing paper. While the intellectual heritage linkage role of citations is probably the dominant consideration when viewing citations as a measure of research impact, one needs to be careful on this point of important contributors to intellectual heritage. In the best of all worlds, only a small fraction of all potential intellectual sources will be and can be acknowledged. Especially in any technical field, there are thousands of papers and other sources which have contributed to the intellectual foundation, as there are thousands of bricks which contribute to the support structure of a building's roof. In particular, there may be sources which are not obvious, at least consciously, to the paper's author. Perhaps a major foundational concept for a paper came from attendance at a seminar or a lunchtime discussion, either of which have escaped the author's memory. Intrinsically, the intellectual attribution process is very incomplete.

Given the finite space allowed in the journals, only a small sampling of the total true intellectual foundation for a paper can be cited, even if all these sources were tangible and identifiable by the author. The selection process used by an author to include a relatively few citations in the bibliography for identifying the intellectual heritage is poorly understood. While some sort of Lotka's law approach is assumed to be at work in selecting only the seminal contributions to the foundation, serious questions exist: what are the selection criteria; what are the cutoff criteria? This uncertainty therefore translates into an undefined role for citations as a measure of intellectual heritage. Some studies (*MacRoberts*, 1996) have attempted to measure the

fraction of intellectual heritage that selected papers included in their bibliographies. While these studies are insightful and useful, the benchmark used (the analyst's perception of what the main intellectual heritage is) is also selective and arbitrary, and limits the utility of such analyses. A more useful approach might be a few case studies where all the references in a sample of published papers are discussed with the authors, and the reasons for inclusion of each reference (and exclusion of other potential references) in the papers are enumerated.

1.3. Citations for tracking research impacts

One critical element of the research management process is identifying and articulating the impact and benefits of research. This helps convince the research sponsors that there has been (or will be) payoff from their research investment, and provides the rationale for continuing the research investment. However, tracking the impacts of research is notoriously difficult. In the process of having impact, research undergoes a transformation to development and engineering, and is effectively camoflouged. Also basic research typically has a multiplicity of impacts in diverse fields. Many of these fields are unfamiliar to the researcher and the sponsor, and therefore any impacts far afield from the researcher's discipline go unrecognized.

For basic research, these latter indirect impacts are an important component of the research's total impact (*Kostoff*, 1994). While the magnitude of these indirect impacts may be small in many (not all) cases, because of the large number of indirect impact pathways, the cumulative effect of all the small indirect impacts resulting from a body of research may be quite large. In fact, in some cases this cumulative effect of indirect impacts could dominate the direct impacts of research (*Kostoff*, 1994).

One largely unutilized role of citations is to serve as a 'radioactive tracer' of research impacts. Citations allow the analyst to track the documented flow and evolution of research over time until the linkages to far downstream products can be identified. Citations allow the different types of impacts to be identified as well. For example, the sponsors of mission-oriented research may want to ascertain whether: 1) certain types of technical disciplines are accessing the research products; 2) certain types of organizations; or specified countries, are utilizing the research products; 3) the research is having its initial direct impact on other basic research or applied research or development. Citations are a documented approach to generating this important diagnostic information.

However, using citations for this diagnostic purpose is much more difficult, complex, and time-consuming than the mainstream application of counting citations for relative impact. The mainstream use of citation counts is algorithm based, and large volumes of data can be processed rapidly to provide copious relative impact results. The tracking application is intrinsically slow and laborious, requiring judgement of the appropriateness

and quality of the impact as well as impact quantity. Because of the potential information available from the tracking application, this is a very fruitful area for future citation research and analysis.

Other positive (and negative) uses of citations can be found in *MacRoberts* (1996) and *Kostoff* (1997a).

1.4. Citations for self-serving purposes

Citations also play other roles, of a less positive (to the advancement of science, anyway) nature. One role is self- aggrandizement, or the ego satisfaction of self-citation for purposes not justified technically. Another role for citations is political. For example, we all know of cases where including citations to journal editors or potential reviewers or 'politically correct' papers will help a paper's chances of being accepted for publication in a specific journal.

Because citations can impact rewards such as promotion/ tenure/ grant consideration, there is a financial self-interest role based on increasing citation volume. This is where 'citation clubs' are formed, and each member cites the other members regularly. Each member has increased citation volume, which eventually translates to more money for each member due to promotions or contracts or other benefits. In addition, there is a potential exclusivity role for citations, whereby they are used mutually among closed groups of researchers to exclude (by sheer volume of citations) competitive concepts which threaten existing mainline infrastructures (see the 'Pied Piper Effect' in Section 2).

2. Citation analysis

2.1. Conclusions from Section 1

Sub-section 1 described some of the many possible uses of citations, including bookmark, intellectual heritage, impact tracker, and self-serving purposes. Since the main published uses of citation analyses tend to focus on absolute and relative measures of impact (and inferred measures of quality), the discussion in this section will concentrated on the applicability of citation analyses as an impact or quality measure. The main message to be derived from sub-section 1 is that there are many reasons for an individual to select particular references for inclusion in a paper, only one of which is the dominant contribution of citations to research impact, significant intellectual heritage. Trying to draw conclusions about the quality or impact of a specific reference based on one particular paper's list of references is akin to solving the inverse problem in science: there may be many solutions; they are not unique; the correct solution cannot be determined without other information. What meaning, then, can be ascribed to the field of citation analysis and the metric of citation counts if the basic unit has such associated uncertainty? More importantly, what is the purpose of using such a metric, and why is its use so widespread?

2.2. Expanded utilization of quantitative measures

While there may be many reasons for the growth and utilization of citation analysis, its expanded use stems (from my perspective) from the evolution of research sponsorship. Technical research has evolved from a rich man's pastime (*Science*, 1998) to industrial support to almost exclusive government support. The approaches used by industry to assess the value of basic research were primarily based on economics. Existing economic tools show that basic research, with its short term costs and long-term high risk payoff horizons, could not be justified as economically cost-effective by most industries. Therefore, since research is viewed by society as a necessity, the support for research has by default almost exclusively shifted to government.

As the U.S. national debt has increased drastically in the last two decades, competition for scarce funds in the Federal arena has increased substantially as well. Basic research, with its long- term payoff horizon, now has to compete strongly with medicare, welfare, and other service provision and development programs. In Europe and Asia, basic research has undergone a similar transformation, with more of a strategic focus to the research.

In this environment of scarce government funds, accountability of all government programs has increased substantially. There are two major characteristics of this increased accountability: more detailed programmatic information is requested by the program assessors, and more quantified information is requested. The upsurge in computer availability over the past decade has enabled large quantities of detailed information to be stored,

tracked, and interpreted, and has driven the request for the large volumes of detailed program information. The request for increased quantitative information also derives from the increased computer capabilities for handling and analyzing large amounts of this type of data. In addition, there is substantial motivation from the assessors to have simple quantitative indicators which could drive the resource allocation process, and substantiate and justify the resource allocation decisions that are generated, rather than use the more complex and expensive and subjective qualitative peer review evaluation processes.

This desire for increased accountability, focused on quantitative measures of research output and impact, counterbalanced by the intrinsic long-term uncertain payoff from research, has produced a dilemma. The simple research outputs, such as published papers and patents, can be easily quantified in the short term. However, they are intermediate measures, not long-tern benefit measures. The quantifiable impacts from research such as societal outcomes or economic payoffs are long-term phenomena and cannot be generated in the short term. Because the research oversight organizations want valid performance metrics applicable to existing research (*Kostoff*, 1997b), the question arises whether credible short term proxies for long-term research impacts and outcomes can be defined.

Citation analyses generate relatively short-term quantifiable items, they have the appearance of short-term research impacts, and are therefore attractive candidates as short-term proxies for research impact and perhaps quality. The real question becomes: what, if anything, do they measure?

2.3. Enhanced value of aggregating citations

The previous section showed that any citation, or group of citations, in a particular paper's bibliography does not provide a unique indicator of positive impact of the cited source on the citing paper. Is there any combination of citations possible which could translate into research impact or quality?

Possibly. Consider the following analogy to gas dynamics. Assume there is a flowing gas with gross velocity V and constant temperature T and pressure P. If one examines a group of molecules in the gas, each member of the group will have a different direction and magnitude to its velocity vector. Thus, the aggregate characteristics of the gas cannot be related to the velocity and 'kinetic temperature' of any one molecule. However, by summing over the velocity distribution functions of large groups of molecules (i.e., taking 'moments' of the velocity distribution function), gross gas properties such as V and P and T can be obtained.

In gas kinetics, one way of viewing each component molecule in its relation to the aggregate is to conceptualize the molecule's velocity vector as consisting of a component with mean velocity V (the aggregate velocity) and a component with random velocity. In the summation process used to derive

aggregated gas properties, the random component is integrated out, leaving only the mean component *V*. Can an analogous model be applied to citation analysis?

I believe it can. Assume that some, if not most, citations reflect intellectual heritage. For any single paper, the citations which reflect intellectual heritage may not be obvious, and of those citations which do reflect intellectual heritage, the dominant or highest priority ones may not be obvious. However, from the nature of the positive and negative reasons for citing shown above, it appears that the main positive reason (intellectual heritage) for citation impact or quality purposes is tied to or reflective of intrinsic technical considerations, and the negative reasons are related to non-technical self-serving individual characteristics. Thus, if we view a paper's bibliography as consisting of a directed (research impact or quality) component related to intellectual heritage and random components related to specific self-interest topics, then for large numbers of citations from many different citing papers, the most significant intellectual heritage (research impact or quality) citations will aggregate and the random author- specific self-serving citations will be scattered and not accumulate.

2.4. Limitations of citations as stand-alone measures of impact

While corroborations of large numbers of citations with other indicators of substantial research impact and quality have shown general agreement, especially with use of large citing and cited universes, there are at least two limitations to this model of citation analysis for stand-alone use as a measure of research impact or quality. First, the reference to intellectual heritage can be positive or negative. A paper could be highly cited because it contributed to the growth of a field, or it could be highly cited because its flaws were obvious to many people, and they wanted to correct the record. Second, there could be systemic biases which affect the aggregate results, one of which I have termed the "Pied Piper Effect" (*Kostoff*, 1997c), and will describe here briefly.

2.5. The Pied Piper Effect

Assume there is a present-day mainstream (characterized by high citations) approach in a specific field of research; for example, the chemical/ radiation/ surgical approach to treating cancer (See *Appendix* 1 for a more detailed example of the "Pied Piper Effect"). Assume that in, say, fifty years a cure for cancer is discovered, and the curative approach has nothing to do with today's mainstream highly-cited research. In fact, assume it turns out that today's highly-cited mainstream approach was completely orthogonal or even antithetical to the correct approach, and that one of the alternative lowly-cited approaches existing today provided the foundation for the eventual cure. Then what meaning can be ascribed to those research papers in cancer today that define the mainstream approach; i.e., they are highly cited for supposedly positive reasons?

In this case, a paper's high citations are a measure of the extent to which the paper's author(s) has persuaded the research community that the research direction contained in his paper is the correct one. The citations are not a measure of the intrinsic correctness of the research direction. In fact, the citations may reflect the desire of a closed research community (the author and the citers) to persuade a larger community (which could include politicians and other resource allocators) that the research direction is the correct one. The citations become the operational mechanism by which the established infrastructure is able to protect its intellectual and capital investments and exclude other competitive approaches which could threaten the integrity of that infrastructure. Citations become the vehicle by which scientific monopoly is established and perpetuated.

This is the "Pied Piper Effect". The large number of citations in the above medical example becomes a measure of the extent of the problem, the extent of the diversion from the correct path, not the extent of progress toward the solution. The "Pied Piper Effect" is a key reason why, especially in the case of revolutionary research, citations and other quantitative measures must be part of and subordinate to a broadly constituted peer review in any credible evaluation and assessment of research impact and quality (*Kostoff*, 1997a).

2.6. Case study of comparative citations

This section ends with a description of a recent short citation study which eventually led to a citing comparison of some Russian/ American papers in different technical fields. The questions raised in interpreting the data highlight a few of the difficulties in attempting to interpret citation results without supplementary information.

In a Text Mining study (*Kostoff*, 1998a) of hypersonic/ supersonic flow over aerodynamic bodies, I generated publication and citation distribution functions for different parameters (authors/ journals/ organizations/ countries). I observed large numbers of authors/ papers/ journals with relatively few citations each, and a few authors/ papers/ journals with large

numbers of citations. I then performed small focused studies to determine the characteristic differences between highly cited and lowly cited papers in hypersonic flow.

Appendix 2 summarizes the results from these focused studies. A key point is that Russian publications tended to populate the poorly cited papers sample, and NASA (U.S.A.) publications tended to populate the highly cited papers sample. To study this Russian/ American difference further, I examined all the papers in the Science Citation Index (SCI) written by the three most prolific Russian authors and the three most prolific American authors in hypersonic/ supersonic flow (names were obtained from the larger Data Mining study). The results, also shown in Appendix 2, were equally striking. Essentially, the Russian papers in this field are not being cited by the larger technical community, or even the Russian technical community.

Because of these findings, I performed another small focused study on the field of near-earth space. I chose this field since we had examined it for a previous Data Mining study (*Kostoff*, 1998b). I selected all English language papers published in 1993 in the SCI (with Russian-Acad-Sci authors only) which contained the word SATELLITE*. I chose Russian-Acad-Sci authors because they were the most prolific according to the larger space Data Mining study.

There were 29 such papers, of which 16 were both relevant to satellites in space and were written by Russian authors only. For each of the 16 papers, I then attempted to identify a paper published by American authors only in 1993 which had at least one reference in common with the Russian paper, and had an approximately similar theme. The SCI allows this capability. I found seven of these pairs; unfortunately, there were not always American papers which met the necessary criteria for pairing with the Russian papers.

Of the 16 relevant Russian papers, 14 had zero cites, one had four cites (two self cites), and one had six cites (two self cites). For the seven pairs of Russian/ American papers, the Russian citation average was 1.4 cites per paper, and the American citation average was about 34 cites per paper (of which about 6.5 were self cites, or about 20%). Also, for these seven pairs, the Russian median was zero cites per paper, and the American median was 37 cites per paper. This is not a large sample, but the differences are so great that I suspect a large sample would give about the same message.

Finally, I performed a small focused study on the field of fullerenes. I selected all English language papers in the SCI published in 1993/ 1994 which contained the phrase CARBON NANOTUBE*. This is one of the 'hottest' areas of fullerene research. There were 131 such papers, all were relevant to the desired topic. I then examined the citation patterns of papers written by Russian authors only and American authors only.

There were 44 papers published by American authors only, and three papers by Russian authors only. The American papers averaged 27.3 cites per

paper, while the Russian papers averaged 6 cites per paper. The American median was 20 cites per paper, while the Russian median was 4 cites per paper. (As an aside, the Japanese papers appeared to very numerous and well cited, followed by the Western European papers).

I may examine other fields, and I may use larger samples, but I seem to be getting a loud and clear message. Whether or not the Russians are prolific in a field in terms of paper production, their works are not getting cited by the larger technical community. Possible explanations are:

1) They could be doing good (citeable) work, and not reporting it;
2) The work reported may be good, but very applied, and not amenable to citing in the literature; i.e., citation is not the appropriate measure of quality or utility or impact in this case;
3) The work reported could be good, but might not be published in the forefront literature, and the technical community therefore might not be very aware of this work.
4) The work could be poor, and the citations pinpoint this.

I have asked perhaps a dozen experts for explanations of these findings, and the number of reasons given approaches the number of experts. This potential diversity of explanations for citation analysis results pinpoints the major operational problem with using these indicators in stand-alone mode.

In the mid-1970s, I led two delegations on Controlled Fusion to the Soviet Union. I visited the Kurchatov Institute in Moscow, and Academgorod near Novosibirsk. Both times, I was impressed by the technical quality of the Russian work in Fusion (both fast-pulsed systems and near-steady state), although there were obvious gaps. At the time, I had the impression that this high technical quality extended to other fields, with obvious exceptions in computers, microelectronics, etc. The present citation results seem to reflect a different level of technical performance than what I thought I had seen in the mid-1970s.

Did I have a misperception then? Had I examined citation performance 20 years ago, would I have arrived at the same conclusions as today? Or has the dissolution of the Soviet Union resulted in a real degradation of their technical performance? Or, are my study approach and groundrules overly limited and not applicable? Or do all of the above explanations and questions have some validity, and point out graphically the deficiencies of trying to use simple quantitative indicators in a stand-alone mode (such as citation counts) to measure extremely complex and sophisticated issues.

2.7. Citation analysis as a warning signal

Perhaps this particular example has shown the value, if any exists, of using quantitative metrics such as citation counts for research quality or impact studies. The quantitative results serve as the 'red flags' or warning lights that a problem may exist; they are the modern day equivalents of the 'canary in the mine' approach to volatile gas detection. However, it was uncertain

exactly what killed the canary decades ago, and it is uncertain today what specific citation counts mean. This is precisely how I use citation studies today; they serve as indicators that further investigation into specific areas is warranted, and they are always accompanied by, and subordinate to, expert analysis/ peer review.

Appendix 1 of Section 8

The Pied Piper Effect: A specific example

A 1995 article in Science purports to identify the Top 10 U.S. Universities in Clinical Medical Research from 1990-1994 (*Science*, 1995). The published papers and citations per paper are ranked in decreasing frequency by medical research institution, and the institutions with the highest frequencies of publicationa and citations are identified as the top universities in clinical medicine research. This Science article crystallizes the problem of using metrics as a gauge of research productivity and, by inference, quality. This statement will be amplified with an illustrative example which questions the linkage between high research output and high research quality. The example focuses on cataracts, but is extrapolateable to other chronic systemic problems as well.

In 1995, I did a literature survey of research papers related to eye cataracts. I examined four years (1991-1994) of abstracts from the *Science Citation Index* (SCI) and the *Social Science Citation Index* (SSCI). Of the many hundreds of abstracts identified, perhaps 99% dealt with different aspects of the surgical treatment of cataracts. Maybe 1% or less dealt with nutritional approaches, and these were mainly vitamin and mineral supplementation for prevention. There were no papers in these peer-reviewed journals dealing with alternative approaches to cataract treatment.

The mainstream medical community views cataracts as an eye problem. The lens degenerates for unknown reasons, in their view, and when it has deteriorated sufficiently, it should be replaced surgically. This approach arises from the paradigm of viewing the eye as a separate component of the total physical system, and the lens replacement becomes equivalent conceptually to replacing a car's windshield when it has become pitted.

An alternative paradigm, subscribed to in part by practitioners such as Leslie Salov, M.D., Gary Todd, M.D., Fereydoon Batmanghelidj, M. D., and Ben Lane, O.D., is that the body experiences chronic systemic problems (deficiencies of various types), and these problems manifest themselves as symptoms in specific organs. For some people, the weak organ is the eye, and the symptom is the cataract. Healing, in this paradigm, consists of identifying and eliminating the systemic deficiencies. Surgically removing the cataract, while improving functioning (at least temporarily), does nothing to address the fundamental problems which are at the foundation of the cataract's presence. Under this alternative paradigm, surgical removal is equivalent to removing the warning light on a car's dashboard when it signifies a problem. In both cases, the long-term consequences of the short-term pallative actions could be very severe.

These alternative approaches never surface in the peer reviewed literature, as the author's survey showed. The journal reviewers (and the funding proposal reviewers as well) are researchers trained along the orthodox paradigms, and

they provide high marks to those papers (and proposals) aligned with the reviewers' backgrounds. In addition, there are institutional and commercial biases which also govern the willingness of the reviewers and editors (and sponsors) to provide positive evaluations of alternative approaches. Thus, the copious papers and citations (and grants) from this component of medical research reflect activity among a closed group whose members subscribe to essentially the same orthodox paradigm. Far from being a measure of quality, the numbers of papers and citations (and projects) from some branches of medical research could be interpreted as a measure of the extent of the problem.

In studies on research evaluation (*Kostoff*, 1995), I differentiate between the two major characteristics of high quality science, doing the *job right* and doing the *right job* (in the best of all worlds, the right job would be done right). The *Science* article (*Science*, 1995) is an example of doing the job right. Once the research target has been selected (paradigm of using the surgical approach to eliminating cataracts), the orthodox medical research community performs an excellent and highly productive effort in finding the best ways to achieve the target. It is analogous to firing a missile very precisely at the wrong target, and is the essence of achieving high precision with low accuracy. However, one can question seriously whether the orthodox medical community is doing the right job (using the right paradigm), and the present closed funding, review, and publication structure effectively precludes innovations which will address the right job.

The *Science* article, and the above comments, illustrate the danger of relying on metrics alone to infer quality from scientific activity. Metrics have an important role in a comprehensive evaluation procedure of research (*Kostoff*, 1997a), but as a stand-alone approach as reflected in the Science article they are subject to misinterpretation.

Appendix 2 of Section 8

Characteristics of highly-cited and poorly-cited papers

To ascertain whether any relationship between highly cited and lowly cited papers and their associated journals and performing organizations could be observed, the characteristics of samples of highly and lowly cited papers were analyzed. The database used to extract the samples was the expanded web version of the SCI. In contrast to the CD-ROM version of the SCI used to obtain the bulk data for this paper, the web version has 60% more journals (~5200), and is more convenient for performing citation analyses (however, the web version in its present incarnation is less convenient than the CD-ROM version for most bulk data analysis, since each record must be downloaded individually). All records in the web version which contained the term HYPERSONIC (a small subset of the supersonic/ hypersonic field) and were published in 1993 were examined.

There were 155 raw 'hits', or records obtained by the query, of which 15 (10%) were not applicable to the topic of hypersonic flow over aerodynamic bodies. Of the remainder, 64 records (46%) had zero citations by other papers; 55 records (39%) received between one and four citations; and 21 records (15%) were cited five or more times by other documents in the expanded SCI, and were viewed as highly cited papers.

Seven of those highly cited papers (33%) were published in the *AIAA Journal* (231-number of papers from database published in journal); three papers in the *Journal of Spacecraft and Rockets* (109); three papers in the *Journal of Fluid Mechanics* (48); and one paper each in a variety of journals which contained fewer papers from the total database. The median journal in the sample contained 48 of the total database papers, as contrasted to the median journal in the total database containing one paper. Since the number of journals which contain n published papers follows approximately a 1/n^4 distribution as was shown in *Kostoff* (1998a), the journals in the highly cited sample are, on average, the very top echelon of the total database journals in terms of numbers of papers published.

In the highly cited paper sample, twelve were from foreign institutions; twelve were from universities; and six were from NASA laboratories. The five most highly cited papers were from universities. The median organization in this sample contributed thirteen papers to the total database, as contrasted to the median organization in the total database contributing one paper. Since the number of papers n contributed by an organization to the total database also follows a 1/n^4 distribution as was shown in *Kostoff* (1998a), the organizations in the highly cited sample are, on average, the very top echelon of the total database organizations in terms of numbers of papers contributed.

The 64 records with zero citations were also examined, albeit from a different perspective. Because the range of citations in the total 140 record

sample was between zero and ten, it was felt that there probably was a quality stratification within the sample group with zero citations, and thus the very poor performers could not be isolated as precisely as the good performers. The following observations were made of the zero cited papers sample.
AIAA Journal contributed 3% of the zero cited papers, as contrasted to 33% of the papers in the highly cited sample; *Journal of Spacecraft and Rockets* – 13% zero cited/ 14% highly cited; *Journal of Fluid Mechanics* – 0% zero cited/ 14% highly cited; HIGH TEMPERATURE - 9% zero cited/ 0% highly cited; *Journal of Aircraft* - 8% zero cited/ 0% highly cited; *PMM Journal of Applied Mathematics and Mechanics* – 6% zero cited; 0% highly cited; *Zeitschrift für Flugwissenschaften und Weltraumforschung* -6% zero cited/ not listed in CD-ROM database. The journals with a high ratio of highly cited papers to zero cited papers tend to emphasize the more fundamental research. The journals with a low ratio of highly cited papers to zero cited papers tend to emphasize the more applied research. The fact that the applied papers are being cited less than the more fundamental papers does not mean they are less useful or of lower quality; they may be of substantial use to developers, who publish much less than researchers, and this more practical use would not be reflected in the present type of bibliometrics study.

Industrial organizations contributed 27% of the zero cited papers, as contrasted to 10% (2 papers) of the highly cited papers (these two highly cited papers were actually one paper split into two sections and published sequentially in the same journal issue); university organizations -33% zero cited; 57% highly cited; NASA - 9% zero cited/ 29% highly cited; American organizations -36% zero cited/ 43% highly cited; European organizations - 25% zero cited/ 38% highly cited; Asian organizations -9% zero cited/ 14% highly cited; Middle Eastern organizations -5% zero cited; 0% highly cited; Russian organizations -23% zero cited; 5% highly cited. This last observation is quite surprising, since two of the top four paper contributing organizations in the total CD-ROM database were Russian.

In summary, this small sample analysis led to the following conclusions for hypersonic flow. Fundamental research papers are more likely to be cited than applied research papers; university papers are more likely to be cited than industry papers; the journals which contain concentrations of highly cited papers are also the core journals in terms of papers published; NASA produced many papers (147 in the total CD-ROM database), and had a substantial fraction of the highly cited papers; Russia produced slightly more papers than NASA (169 in the total CD-ROM database), and had almost no highly cited papers.

The NASA/ Russia citation differential led to another short study which examined American/ Russian differentials in supersonic/ hypersonic flow citations. Two groups of papers were generated. The first group consisted of all papers (from the web version of the SCI) published in 1993/ 1994 by the three most prolific supersonic/ hypersonic flow Russian authors identified in *Kostoff* (1998a); the second group included all papers by the three most

prolific supersonic/ hypersonic flow American authors from Kostoff [1998a]. There were 12 papers in the first (Russian) group, and 36 papers in the second (American) group. All papers related to supersonic/ hypersonic flow. The citations received by all these papers were examined.

Of the twelve Russian papers, nine received zero cites, two received one cite each, and one received three cites. The average cites per paper is 0.4. All of the five total cites were self- cites (There is nothing intrinsically wrong with self cites; in those cases where the author has done the pioneering work in the field, self-cites are most appropriate. However, when all cites are self-cites, then the true impact of the paper on the larger scientific community must be called into question).

Of the 36 American papers, seven received zero cites. The total number of citations received was 106, of which 56 were self cites. The average cites per paper is three. While all these citation numbers reported are quite small, reflecting the low level of effort in this technical field, there is obviously a systemic difference between the citations received by the Russian and American papers. Whether these differences extend beyond supersonic/ hypersonic flow to other topical areas is an interesting question.

There are two crucial pieces of data missing from these two short studies (and from most bibliometrics analyses) which prevent harder conclusions about quality and value to be drawn. The amount of research effort represented by each paper is unknown to the analyst, and the eventual use of the results from each paper is unknown to the analyst. Thus, the number of highly cited papers per dollar of research investment (or some similar research efficiency metric), probably a better measure of value than pure numbers of papers or highly cited papers, cannot be stated. Also, the quality of the eventual hypersonic vehicles that resulted from the papers' research, probably a better measure than numbers of cited papers, was not tracked and cannot be stated. In addition, the papers in these two short studies were not read in detail independently by hypersonic flow experts, and thus their quality could not be gauged independently from another perspective and correlated to the citation results.

REFERENCES FOR SECTION 8

KOSTOFF, R. N. (1994), Research impact quantification, *R&D Management,* 24:3, July.

KOSTOFF, R. N. (1995). Federal research impact assessment: Axioms, approaches, applications, *Scientometrics,* 34:2.

KOSTOFF, R. N. (1997a), *The Handbook of Research Impact Assessment,* Seventh Edition. DTIC Report Number ADA296021. Also, available at http://www.dtic.mil/dtic/kostoff/index.html

KOSTOFF, R. N. (1997b), Peer review: The appropriate GPRA metric for research, *Science,* Vol. 277. 1 August. p. 651-652.

KOSTOFF, R. N. (1997c), Use and misuse of metrics in research evaluation, *Science and Engineering Ethics,* 3:2.

KOSTOFF, R. N. (1998a), Hypersonic and supersonic flow roadmaps using bibliometrics and database tomography, (submitted for publication).

KOSTOFF, R. N. (1998b), Database tomography for technical intelligence: A roadmap of the near-earth space science and technology literature, *Information Processing and Management,* (accepted for publication).

LEYDESDORFF, L. (1998), Theories of citation, *Scientometrics* (this issue).

MACROBERTS, M. and MACROBERTS, B. (1996), Problems of citation analysis, *Scientometrics*, 36 (3) 435.

Top 10 U. S. universities in clinical medicine research, 1990-1994, *Science*,. Vol. 269. 1 September, 1995. p. 1223.

Scientists who fund themselves, *Science,* Vol. 279. 9 January, 1998, p. 179.

Section 9. Citation Mining for Research Impact Identification.

(based on Kostoff, R. N., Del Rio, J. A., García, E. O., Ramírez, A. M., and Humenik, J. A. "Citation Mining: Integrating Text Mining and Bibliometrics for Research User Profiling". JASIST. 52:13. 1148-1156. 52:13. November 2001.)

I. OVERVIEW

Background: Identifying the users and impact of research is important for research performers, managers, evaluators, and sponsors. It is important to know whether the audience reached is the audience desired. It is useful to understand the technical characteristics of the other research/ development/ applications impacted by the originating research, and to understand other characteristics (names, organizations, countries) of the users impacted by the research. Because of the many indirect pathways through which fundamental research can impact applications, identifying the user audience and the research impacts can be very complex and time consuming.

Objective: The purpose of this section is to describe a novel approach for identifying the pathways through which research can impact other research, technology development, and applications, and to identify the technical and infrastructure characteristics of the user population.

Approach: A novel literature-based approach was developed to identify the user community and its characteristics. The research performed is characterized by one or more papers accessed by the Science Citation Index (SCI) database, since the SCI's citation-based structure enables the capability to perform citation studies easily. The user community is characterized by the papers in the SCI that cite the original research papers, and that cite the succeeding generations of these papers as well. Text mining is performed on the citing papers to identify the technical areas impacted by the research, the relationships among these technical areas, and relationships among the technical areas and the infrastructure (authors, journals, organizations). A key component of text mining, concept clustering, was used to provide both a taxonomy of the citing papers' technical themes and further technical insights based on theme relationships arising from the grouping process. Bibliometrics is performed on the citing papers to profile the user characteristics. Citation Mining, this integration of citation bibliometrics and text mining, is applied to the ~300 first generation citing papers of a fundamental physics paper on the dynamics of vibrating sand-piles.

Results: Most of the ~300 citing papers were basic research whose main themes were aligned with those of the cited paper. However, about twenty percent of the citing papers were research or development in other disciplines, or development within the same discipline. The text mining alone identified the intra-discipline applications and extra-discipline impacts and applications; this was confirmed by detailed reading of the ~300 abstracts.

Conclusions: The combination of citation bibliometrics and text mining provides a synergy unavailable with each approach taken independently. Furthermore, text mining is a REQUIREMENT for a feasible comprehensive research impact determination. The integrated multi-generation citation analysis required for broad research impact determination of highly cited papers will produce thousands or tens or hundreds of thousands of citing paper Abstracts. Text mining allows the impacts of research on advanced development categories and/ or extra-discipline categories to be obtained without having to read all these citing paper Abstracts. The multi-field bibliometrics provide multiple documented perspectives on the users of the research, and indicate whether the documented audience reached is the desired target audience.

II. BACKGROUND

Identification of diverse research impacts is important to research managers, evaluators, and sponsors, and ultimately to performers. They are interested in the types of people and organizations citing the research outputs, and whether the citing audience is the target audience. Also, they are interested in whether the development categories and technical disciplines impacted by the research outputs are the desired targets. Since fundamental research can evolve along myriad paths, tracking diverse impacts becomes complex.

Presently, there are three generic approaches to tracking the impact of research: qualitative, semi-quantitative, and quantitative (Kostoff, 1997). Qualitative approaches are variants of peer review. Panels of experts are assembled, and impacts are identified based on the participants' knowledge, and usually personal experiences. The results are usually long on subjectivity, and short on independent documentation.

Semi-quantitative approaches are probably the most widely used for tracking impact (Kostoff, 1994). They include retrospective studies such as Hindsight (DOD, 1969) and Traces (IITRI, 1968), and various types of

research sponsor accomplishment books such as those from DOE (DOE, 1983, 1986) and DARPA (IDA, 1991). A detailed treatment is contained in (Kostoff, 1997). Semi-quantitative approaches tend to be grounded in corporate memory of the participants, although some studies (Narin, 1989) follow the citation trail for supplementation. Their focus is detailed examination of a few high impact cases, rather than a wide-scale identification of many diverse impacts. As in the peer review approach, semi-quantitative approaches also have a high subjective component.

Quantitative approaches are also widely used for impact tracking (Kostoff, 1994). They tend to be divided between economic methods such as cost-benefit and internal rate-of-return (Averch, 1994; Tassey, 1999), and S&T indicators such as publications and patents (Narin, 1994), and their citations. They are the most objective of the three generic methods for tracking and quantifying research impact. However, many assumptions related to cost and benefit allocation are required for the economic studies (Kostoff, 1997). Additionally, many assumptions are required to accept correlation between numerical indicator values and degree of impact.

Thus, one of the gaps of all these impact tracking techniques is objective identification of the full scope of impacts produced by the research. These impacts include both the directly identifiable research impacts and the indirect impacts. For that fraction of performed research that is documented in the technical literature, tracking of direct and indirect research impacts on intermediate and final useful products becomes possible through tracking of generations of citations to the original research. If this wide scale impact information were obtained, then the in-depth studies performed by the semi-quantitative methods could cover an expanded range, or the roadmap of impacts could be presented as a self-contained valuable finding.

Even though the premier database for citation tracking, the Science Citation Index (SCI), contains a number of data fields abstracted from the full-text published papers, past citation-based studies using the SCI have focused almost exclusively on citation counts as an impact metric. Reviews of these citation studies can be found in (De Solla price, 1986; Braun, 1987; Egghe, 1990). The potential impact of citation counts on decision-making is small, since the information content of citation counts alone is very limited. However, these citing records contain a wealth of information in their two main categories of diverse fields. The non-free-text fields, such as Author, Journal, Address, etc, describe the infrastructure characteristics of the citing

community. The free-text fields, such as Title, Abstract, and Keywords (Keywords is not strictly a free-text field, but has sufficient technical characteristics to be included in this grouping), describe the technical characteristics of the impacted research, development, and applications areas.

Use of the SCI non-free-text fields for citing paper bibliometric analysis has been published on a very sporadic basis, and typically only for one or two data fields (Steele, 2000; Herring, 1999; Davidse, 1997). The focus of most of these studies has been on relating citations or citation rates to the few field variables examined. There do not appear to have been any citation studies performed for the specific purpose of user population profiling, where many of the available fields are examined in an integrated manner.

Use of the SCI free-text fields for coupled trans-citation citing paper/ cited paper text mining analysis has not been published, although text mining studies of SCI and other database free-text fields have been reported (e.g., Kostoff, 2000a, 2000b).

III. OBJECTIVES

The objectives of the present section are:

i) Demonstrate the feasibility of tracking the myriad impacts of research on other research, development, and applications, using the technical literature.
ii) Demonstrate the feasibility of identifying a broad range of research product user characteristics, using the technical literature.
iii) Relate thematic characteristics of citing papers to their cited papers.

IV. APPROACH

The present section describes a novel process, Citation Mining, that uses the best features of citation bibliometrics and text mining to track and document the impact of basic research on the larger R&D community across many generations. In Citation Mining, text mining (Kostoff, 2000a, 2000b; Losiewicz, 2000) of the cited and citing papers (trans-citation) supplements the information derived from the semi-structured field bibliometric analyses. Text mining illuminates the trans-citation thematic relationships, and provides insights of knowledge diffusion to other intra-discipline research,

advanced intra-discipline development, and extra-discipline research and development. The addition of text mining to citation bibliometrics makes feasible the large-scale multi-generation citation studies that are necessary to display the full impacts of research.

A proof-of-principle demonstration of Citation Mining for user population profiling and research impact was performed on four sets of cited papers. The papers were selected based on the authors' technical interests, rather than a random representative sample. It was desired to have one group of papers representative of basic research, and another group representative of applied research. Two of the sets were selected Mexican and U. S. applied photo-voltaic research papers, and two of the sets were selected British and U. S. fundamental vibrating sand-pile research papers. The complete detailed results are reported in Del Rio (2000).

The present section focuses on the trans-citation coupled citing paper/ cited paper text mining results for one of the sets, a highly cited U. S. vibrating sand-pile paper (Jaeger, 1992). Vibrating sand-piles are important in their own right, since they model the behavior of granular systems used in agriculture (seeds, grains), geology (avalanches, soil mechanics), construction (gravel, sand), and manufacturing (powders, lubricants, sand-blasting). The underlying phenomena exhibited in their static and dynamic states can be found in many disparate applications, such as fusion confinement, geological formations, self-assembly of materials, thin film structure ordering, shock-wave statistics, and crowded airspace. Statistically, the sand-pile paper selected has sufficient citing papers for adequate text mining statistics. It covers an exciting area of physics research, and its technical sub-themes have potential for extrapolation to other technical disciplines.

The analyses performed were of two types: bibliometrics and text mining, where the text mining was subdivided into two components, phrase frequency analysis and phrase clustering analysis. These different types of analyses are described in the following sections.

IV-A. Bibliometrics Analysis

The citing paper summaries (records) were retrieved from the SCI. Analyses of the different non-free-text fields in each record were performed, to identify the infrastructure characteristics of the citing papers (authors, journals, institutions, countries, technical disciplines, etc). The detailed

analysis methodologies and extensive results are described in Del Rio (2000).

IV-B. Phrase Frequency Analysis

The purpose of the phrase frequency analysis was to manually generate a taxonomy (technical category classification scheme) of the database, from the quantified technical phrases extracted from the free-text record fields. To generate the database, the citing papers' Abstracts were aggregated. Computational linguistics analyses were then performed on the aggregate. Technical phrases were extracted using the Database Tomography process (Kostoff, 2000a, 2000b; Losiewicz, 2000). An algorithm extracted all single, adjacent double, and adjacent triple word phrases from the text, and recorded the occurrence frequency of each phrase. While phrases containing trivial/ stop words at their beginning or end were eliminated by the algorithm, extensive manual processing was required to eliminate the low technical content phrases. Then, a taxonomy of technical sub-categories was generated by manually grouping these phrases into cohesive categories. Intra-discipline applications, and extra-discipline impacts and applications were identified from visual inspection of the phrases.

IV-C. Phrase Clustering Analysis

The purpose of the phrase clustering analysis was to generate taxonomies of the database semi-automatically, again from the quantified technical phrases extracted from the free-text record fields. The clustering analysis further used quantified information about the relationships among the phrases from co-occurrence data (the number of times phrases occur together in some bounded domain). The clustering analyses results complemented those from the phrase frequency analyses, and offered added perspectives on the thematic structure of the database.

After the phrase frequency analyses were completed, co-occurrence matrices of Abstract words and phrases (each matrix element Mij is the number of times phrase or word i occurs in the same record Abstract as phrase or word j) were generated using the TechOasis phrase extraction and matrix generation software. As in the phrase frequency analysis, the phrases extracted by the TechOasis natural language processor required detailed manual examination, to eliminate the low technical content phrases. The co-occurrence matrices were input to the WINSTAT statistical clustering software, where clusters (groups of related phrases based on co-occurrence

frequencies) based on both single words and multi-word phrases were generated.

Two types of clustering were performed, high and low level. The high level clustering used only the highest frequency technical phrases, and resulted in broad category descriptions. The low level clustering used low frequency phrases related to selected high frequency phrases, and resulted in more detailed descriptions of the contents of each broad category.

IV-C-1. High Level Clustering

The TechOasis phrase extraction from the citing Abstracts produced two types of lists. One list contained all single words (minus those filtered with a stop word list), and the other list contained similarly filtered phrases, both single and multi-word. Both lists required further manual clean-up, to insure that relatively high technical content material remained. The highest frequency items from each list were input separately to the TechOasis matrix generator, and two co-occurrence matrices, and resulting factor matrices, were generated.

The co-occurrence matrices were copied to an Excel file, and the matrix elements were non-dimensionalized. To generate clusters defining an overall taxonomy category structure for the citing papers, the Mutual Information Index was used as the dimensionless quantity. This indicator, the ratio of: the co-occurrence frequency between two phrases (i,j) squared (Cij^2) to the product of the phrase occurrence frequencies (Ci*Cj), incorporates the co-occurrence of each phrase relative to its occurrence in the total text. The co-occurrence matrix row and column headings are arranged in order of decreasing frequency, with the highest frequency phrase occurring at the matrix origin. Based on the intrinsic nature of word and phrase frequencies, the row and column heading frequencies decrease rapidly with distance from the matrix origin. With increasing distance from the origin, the matrix becomes more and more sparse, although the phrases themselves have higher but more focused technical content. In parallel, the Mutual Information Index's values decrease rapidly as the distance from the matrix origin increases. Thus, the Mutual Information Index is useful for relating the highest frequency terms only, and for providing the top-level structural description of the taxonomy categories.

IV-C-2. Low Level Clustering

To obtain a more detailed technical understanding of the clusters and their contents, the lower frequency phrases in each cluster need to be identified. A different matrix element non-dimensional quantity is required, one whose magnitudes remain relatively invariant to distance from the matrix origin. In addition, a different approach for clustering the low frequency phrases in the sparse matrix regions is required, one that relates the very detailed low frequency phrases to the more general high frequency phrases that define the cluster structure. In this way, the low frequency phrases can be placed in their appropriate cluster taxonomy categories.

The method chosen to identify the lower frequency phrases is as follows. Start with the cluster taxonomy structure defined by grouping the higher frequency phrases using the Average Neighbor agglomoration technique and the Mutual Information Index. Then, for each high frequency phrase in each cluster, find all phrases whose value of the Inclusion Index Ii exceeds some threshold. Ii is the ratio of Cij to Ci (the frequency of occurrence of phrase i in the total text), where phrase i has the lower frequency of the matrix element pair (i,j). A threshold value of 0.5 for Ii was used. The resultant lower frequency phrases identified by this method will occur rarely in the text, but when they do occur, they will be in close physical (and thematic) proximity to the higher frequency phrases.

V. RESULTS

V-A. Phrase Frequency Analysis

The Abstract of the highly cited vibrating sand-pile paper (Jaeger, 1992) is shown in Figure 1.

FIGURE 1 – CITED PAPER ABSTRACT

Granular materials display a variety of behaviors that are in many ways different from those of other substances. They cannot be easily classified as either solids or liquids. This has prompted the generation of analogies between the physics found in a simple sandpile and that found in complicated microscopic systems, such as flux motion in superconductors or spin glasses. Recently, the unusual behavior of granular systems has led to a number of new theories and to a new era of experimentation on granular systems.

This paper had ~300 citing papers listed in the SCI, as of mid-CY2000. The highest frequency single, adjacent double, and adjacent triple word phrases from the aggregate citing papers aligned with the central themes of the cited paper can be represented by the following generic taxonomy: Theory/ modeling; Experiments/ measurements/ instruments/ variables; Structure/ Properties; Phenomena.

There were hundreds of technical phrases in each taxonomy category, and the authors selected those judged representative of each category for the purposes of illustration. Those representative phrases (*Underlined*) aligned with the central themes of the aggregate citing papers offer the following intra-discipline portrait of the citing aggregate. These papers reflect a balanced theoretical/ modeling effort (Molecular Dynamics Simulations, Monte Carlo Simulations, Kinetic Theory) and experimental effort (Magnetic Resonance Imaging, Charge Coupled Device Camera) targeted at studying the motions of granular particles. The papers focus on examination of the structure(s) and properties of vibrating sand-piles (Angle Of Repose, Coefficient Of Restitution), and the intrinsic phenomena of these collective systems (Collisions Between Particles, Fractional Brownian Motion), with emphasis on segregation (Size Segregation, Axial Segregation, Radial Segregation), relaxation (Relaxation Dynamics, Relaxation Time Tau), avalanching (Avalanch Durations, Avalanch Size), fluidization (Onset of Fluidization, Formation of Convection Cells), and collective behaviors (Collective Particle Motion, Self-Organized Criticality).

While the citing paper phrases mainly reflected emphasis on studies of granular piles, the phenomenological results and insights on segregation, relaxation, fluidization, avalanching, and collective behavior were extrapolated to some extra-discipline applications. These include (sample category abbreviated record Titles follow the phrases):

Category	**Phrases**	**Titles**
geological formations and processes	(Earthquake*, Rock Avalanches, Carbonate Turbudite Deposition),	* *Sedimentary evolution of the early Paleocene deep-water Gulf of Biscay* * *A fragmentation-spreading model for long-runout rock avalanches*
soil mechanics	(Soil Mechanics, Hillslope Gradient),	**Evidence for nonlinear, diffusive sediment transport on hillslopes and implications for landscape morphology* **Analysis of vertical projectile penetration*

		in granular soils
industrial applications	(Screw Feeder*, Industrial),	**Precision dosing of powders by vibratory and screw feeders*
interacting object dynamics	(Traffic Congestion, War Game*).	**Study on crowded two-dimensional airspace - Self-organized criticality* **Derivation and empirical validation of a refined traffic flow*
materials	(Rheolog*, Untwinned Single Crystals, Chemical Shift Tensors),	**Vortex avalanches at one thousandth the superconducting transition temperature* **Mesoscale self-assembly of hexagonal plates using lateral capillary forces*
films	(Molecular Fluids, Adsorbed Polymer Layers),	**A model for the static friction behaviour of nanolubricated contacts* **Spontaneous formation of ordered structures in thin films of metals*
multi-phase systems	(Flow Immunosensors, Fluidized Bed*),	**Advances in flow displacement immunoassay design* **Rheophysical classification of concentrated suspensions and granular pastes* **From bubbles to clusters in fluidized beds*
gas dynamics	(Gas Flow, Shock Waves, Shock Front),	**Statistics of shock waves in a two-dimensional granular flow* **Scale invariant correlations in a driven dissipative gas*
micro particles	(Pollen Exine Sculpturing, Molecular, Spinule)	**The effects of genotype and ploidy level on pollen surface sculpturing in maize*
and microscale cooperative effects	(Tokamak, Plasma*, Lattice Gas).	**Sandpiles, silos and tokamak phenomenology: a brief review* **Logarithmic relaxations in a random-field lattice gas subject to gravity*

To validate the text mining results, each of the ~300 citing paper Abstracts was read by the first two authors, and those Abstracts reflecting applications and extra-discipline impacts were identified. *All of the applications and extra-discipline papers identified from reading the Abstracts could be identified/ retrieved from examination of the anomalous text mining-derived phrases with a threshold frequency of two.* The applications taxonomy of

the previous section was validated using this Abstract reading and manual classification process, and judged to be a reasonable classification of the applications and extra-discipline impacts. Identification of the applications and extra-discipline impacts most unrelated to the main themes of the cited paper was easiest because of the highly anomalous nature of their representative phrases. Identification of the intra-discipline applications was the most difficult, since the phraseology used was similar to that of the cited paper themes.

The importance of this result should be emphasized. A complete citation impact study will typically involve multiple generations of citations. For a citation impact study that involves large numbers of first-generation citing papers and/ or large numbers of succeeding generation citing papers, reading each citing paper Abstract to identify applications paper characteristics becomes infeasible. For example, the ~300 citing papers of the sand-pile paper were themselves cited by ~3600 papers. If and when full-text becomes available for citation analysis, the time to read each paper will increase by an order of magnitude.

Use of text mining capabilities, such as computational linguistics, allows only those applications and extra-discipline papers of interest to be identified, and the requisite information could then be obtained from reading the Abstracts. In addition, the computational linguistics provides a structure and categorization of these myriad applications, allowing the larger context of application themes to be displayed and understood.

The citing papers representing categories of development and disciplines aligned and non-aligned with those of the cited paper are shown in the matrix of Figure 2.

FIGURE 2 – DEVELOPMENT CATEGORY AND CITED PAPER THEME ALIGNMENT OF CITING PAPERS

TECH DEV	33									
TECH DEV	32			1						
TECH DEV	31									
APPL RES	23						1		1	1
APPL RES	22			1				3		
APPL RES	21				1		1			

BAS RES	13	1	2	2	2	2	3		1	
BAS RES	12		2	3	6	4	10	8	10	1
BAS RES	11	3	23	28	27	43	43	30	33	4
		1992	1993	1994	1995	1996	1997	1998	1999	2000

TIME

CODE: MATRIX ELEMENT IS NUMBER OF PAPERS

In Figure 2, the abscissa represents time. The ordinate, in the second column from the left, is a two-character tensor quantity. The first number represents the level of development characterized by the citing paper (1=basic research; 2=applied research; 3=advanced development/ applications), and the second number represents the degree of alignment between the main themes of the citing and cited papers (1=strong alignment; 2=partial alignment; 3=little alignment). Each matrix element represents the number of citing papers in each of the nine categories.

There are three interesting features on Figure 2. First, the tail of total annual citation counts is very long, and shows little sign of abating. This is one characteristic feature of a seminal paper.

Second, the fraction of extra-discipline basic research citing papers to total citing papers ranges from about 15-25% annually, with no latency period evident. This lag-free extra-disciplinary diffusion may have been due to the combination of intrinsic broad-based applicability of the subject matter and publication of the paper in a high-circulation science journal with very broad-based readership.

Third, a four-year latency period exists prior to the emergence of the higher development category citing papers. This correlates with the results from the bibliometrics component. From the present study, it is not possible to differentiate the reasons for this important result. The latency could have been due to the inability of the technology community to *immediately* recognize the potential applications of the science. Or, it could have been due to the information remaining in the basic research journals, and not reaching the applications community. Or, the time that an application needs to be developed in this discipline is of the order of four years. Thus, the basic science publication feature that may have contributed heavily to extra-

discipline citations may also have limited higher development category citations for the latency period.

Finally, the alignment of the citing journal theme to the main theme of the cited paper was estimated for all citing papers. In essentially all cases, the citing paper theme could be subsumed within the citing journal theme. However, given the breadth of most journal themes, this result had minimal information content (e.g., citing paper X was published in a Physics journal vs. a Materials journal).

In Davidse's study of Physics papers citations (Davidse, 1997), a key metric used in cross-disciplinary citations/ impacts was the distinction between Physics and non-Physics papers. It was implicitly assumed that the flow from Physics to non-Physics papers was analogous to the flow from basic to applied. While it may be true for some cases, Figure 2 (and other unpublished studies) shows that most extra-discipline flows in the present study were from basic physics research to basic research in the other disciplines. *Here, it is important to emphasize that a seminal idea gives new possible interpretations in many other disciplines.*

Davidse used journal themes (based on the SCI journal classification taxonomy) as a proxy for citing paper themes, with the level of resolution being at the gross discipline description, at best. There are many multi-discipline journals (e.g., Science, Nature, etc) that render a thematic distinction impossible. Davidse's approach required such a computerized proxy representation, since tens of thousands of citing papers were analyzed.

In contrast, the present section's approach of identifying impact themes through text mining allows a much more detailed and informative picture of the impact of research to be obtained. It represents the difference between stating that a "Physics paper impacted Geology research" and a "paper focused on sand-pile avalanches for surface smoothing impacted analyses of steep hill-slope landslides".

V-B. Phrase Clustering Analysis

V-B-1. High Level Clustering

For illustrative purposes, a sample truncated co-occurrence matrix based on phrases from the ~300 citing Abstracts is shown on Figure 3.

FIGURE 3

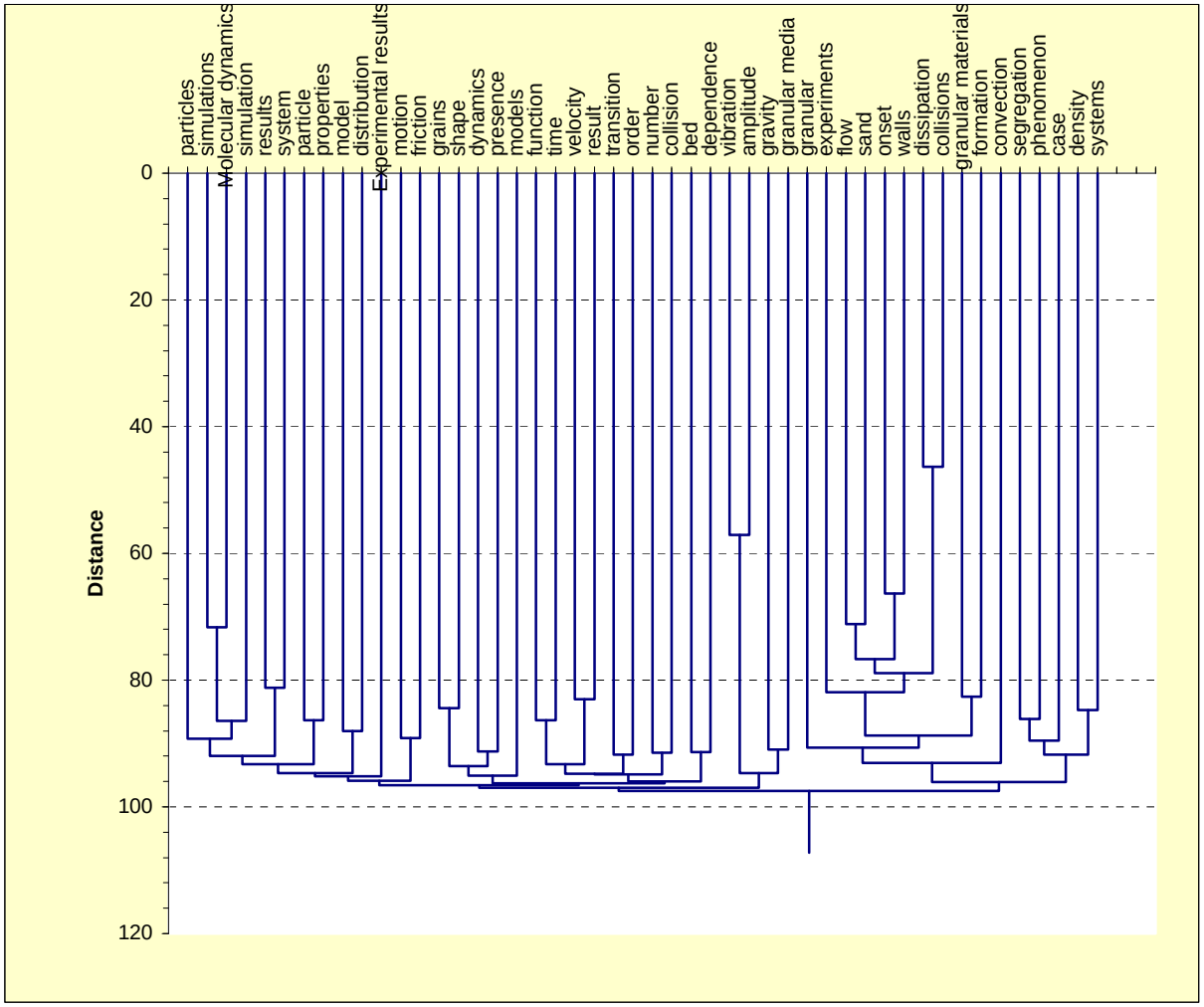
# Records	1 particles	granular	results	system	Experime	granular	materials	grains	Flow	dynamics	motion	simulation	function	number	formation	segregatio
45 particles	45	10	5	8	8	7	8	6	4	7	11	2	4	2	2	
45 granular	10	45	5	4	8	4	8	8	4	6	3		4	5	1	
40 results	5	5	40	12	2	5	8	4	1	4	10	4	3	2	3	
39 system	8	4	12	39	4	1	9	2	9	6	10	3	6	3	2	
37 experiments	8	8	2	4	37	10	6	12	6	9	3	1	3	6	4	
37 granular materials	7	4	5	1	10	37	6	7	3	6	3	1	5	8	6	
37 grains	8	8	8	9	6	6	37	6	5	5	4	4	6	3	6	
34 flow	6	8	4	2	12	7	6	34	5	7	2	4	2	5	4	
33 dynamics	4	4	1	9	6	3	5	5	33	3	4	3	3	2	3	
33 motion	7	6	4	6	9	6	5	7	3	33	3	3	4	4	1	
28 simulations	11	3	10	10	3	3	4	2	4	3	28	2	2	1	4	
25 function	2		4	3	1	1	4	4	3	3	2	25	3	1	2	
21 number	4	4	3	6	3	5	6	2	3	4	2	3	21	1	1	
20 formation	2	5	2	3	6	8	3	5	2	4	1	1	1	20	2	
20 segregation	2	1	3	2	4	6	6	4	3	1	4	2	1	2	20	

In the final data analysis, a clustering of the 153 highest frequency technical content phrases in the matrix rows was then performed using the Excel add-in statistical package WINSTAT. A particularly helpful output for each clustering run was the dendogram, a tree-like diagram showing the structural branches that define the clusters. Figure 4 is one dendogram based on the 48 highest frequency phrases (for illustration purposes only). The abscissa contains the phrases that are clustered. The ordinate is a distance metric. The smaller the distance at which phrases, or phrase groups, are clustered, the closer is the connection between the phrases.

Thus, the first phrases combined are DISSIPATION and COLLISIONS, followed by VIBRATION and AMPLITUDE. At some later time, the VIBRATION-AMPLITUDE combination is grouped with the GRAVITY-GRANULAR MEDIA combination to form the next hierarchical level grouping, and so on. For the 48 phrases selected, the top hierarchical level

consists of two clusters. On Figure 4, one cluster is bounded by the phrases PARTICLES-GRANULAR MEDIA, and the other is bounded by the phrases GRANULAR-SYSTEMS.

FIGURE 4



Many agglomoration techniques were tested; the Average Neighbor method appeared to provide reasonably consistent good results. Analyses were performed of the numerous cluster options that were produced. The following is the top-level cluster description that represented the results of the phrase and word lists clustering best, as well as the factor matrix clustering from the TechOasis results.

The highest level categorization based on the highest frequency 153 phrases produced three distinct clusters: Structure/ Properties, Flow-Based Experiments, Modeling and Simulation. In the description of each cluster that follows, phrases that appeared within the clusters will be capitalized.

1) Structure/ Properties

This cluster contained MIXTURES of LARGE GRAINS and SMALL GRAINS, with STRATIFICATION along ALTERNATING LAYERS based on SIZE SEGREGATION and grain SHAPE and GEOMETRICAL PROFILE. The MIXTURE forms a PILE with an ANGLE of REPOSE. When the ANGLE of REPOSE is LARGER than a critical ANGLE, DYNAMICAL PROCESSES produce AVALANCHES, resulting in SURFACE FLOW within THIN LAYERS.

2) Flow-Based Experiments

This cluster contained EXPERIMENTS examining GRANULAR and SAND FLOW, The dependence of ENERGY DISSIPATION, due to COLLISIONS, on PACKING DENSITY was a focal area. The INFLUENCE of PIPE WALLS and PLATES on the SHEAR-driven VELOCITY and DENSITY PROFILES was studied, as well as ONSET of FLUIDIZATION and CONVECTIVE FLOW with its attendant FORMATION of CONVECTION CELLS.

3) Modeling and Simulation

This cluster contained MODELS and NUMERICAL SIMULATIONS based on EXPERIMENTAL RESULTS, OBSERVATIONS, MEASUREMENTS, and DATA. The SIMULATION METHODS MODELED the CHARACTERISTICS of DYNAMIC EVOLUTION from INITIAL CONDITIONS to STEADY STATE. A strong focal area was the CHARACTERISTICS of POWER SPECTRUM POWER LAW DISTRIBUTIONS, and their ROLE in the DYNAMIC EVOLUTION from INITIAL INSTABILITY to CRITICALITY. Sound PROPAGATION, especially its relation to DEPTH and PRESSURE, as a function of TIME and VIBRATION FREQUENCY, AMPLITUDE, and ACCELERATION is modeled with the statistical mechanics concepts of GRANULAR TEMPERATURE through KINETIC THEORY. Additionally, GRAVITY and VIBRATIONS are PHENOMENA used in the EQUATIONS to model the COMPACTION of GRANULAR MEDIA.

V-B-2. Low Level Clustering

Four types of results were obtained with the lower frequency phrases. Many of the lower frequency phrases were closely associated with one higher frequency phrase only; most lower frequency phrases were closely associated with one of the three clusters only; a few lower frequency phrases were associated with more than one cluster; and only a majority of the lower frequency phrases that related to applications or other disciplines were uniquely related to a single cluster. Sample relationships from each of these four types follow.

a) Lower Frequency Phrases Unique to One Higher Frequency Phrase
(High Frequency Phrase: Low Frequency Phrases)

REPOSE: VIBRATIONAL ACCELERATIONAL AMPLITUDE; STRATIFICATION: FACETED GRAINS; FLOW: VERTICAL GLASS PIPE, KINEMATIC SIEVING; COLLISIONS: LONG-RANGE CORRELATIONS; MODEL: COUPLED NONLINEAR STOCHASTIC EQUATIONS, SELF-ORGANIZED CRITICALITY; SIMULATION: DISCRETE ELEMENT METHOD; RELAXATION: STRONG SPATIAL CLUSTERING.

The phrases in this category, on average, tend to be longer and more detailed/ specific than the phrases in any of the other categories. They also tend to be the lowest frequency phrases, and their length and detail characteristics are consonant with the very lowest frequency phrases.

b) Lower Frequency Phrases Unique to One Cluster (Cluster High Frequency Phrases: Low Frequency Phrase)

LARGE GRAINS, SMALL GRAINS, REPOSE, STRATIFICATION: ALTERNATING LAYERS; COLLISIONS, CONVECTION CELLS, DISSIPATION EXPERIMENTS, FLOW, PACKING, VELOCITY PROFILES: ONSET OF FLUIDIZATION; DYNAMICS, RELAXATION: CONFIGURATIONAL ENTROPY; MODEL, SIMULATIONS: MKDV EQUATION

The low frequency phrases associated uniquely with the flow-based experiments cluster tended to be associated with the largest number of high frequency phrases, whereas the low frequency phrases associated uniquely with the modeling and simulation cluster tended to be associated with the smallest number of high frequency phrases. This reflects the more closely-knit nature of the flow-based experiments cluster relative to the more diverse

nature of the modeling and simulation cluster, and was confirmed by examining all the high frequency phrases in each cluster.

c) Low Frequency Phrases Shared by All Three Clusters (High Frequency Phrases: Low Frequency Phrase)

POWER LAW, EXPERIMENTS, AVALANCHE: AVALANCHE DURATIONS; SIMULATIONS, EXPERIMENTS, STRATIFICATION: CONTACT NETWORK; DYNAMICS, ONSET, AVALANCHE: TOP LAYER; MODEL, FLOW, STRATIFICATION: STATIC GRAINS

As a general rule, the low frequency phrases in this category tend to be relatively generic, at least compared to phrases in the other three categories.

D) Low Frequency Phrases from Applications or other Disciplines (High Frequency Phrase(s): Low Frequency Phrase)

DENSITY WAVES: TRAFFIC FLOW; MODEL: AIR TRAFFIC; MODEL: CELL PELLETS; DYNAMICS, MODEL: DUNES; DYNAMICS, FLOW: IMMUNOSENSORS; MODEL, FLOW, AVALANCHES: GEOLOGICAL; MODEL, SIMULATION: WAR GAME; MODEL, DISSIPATION: VISCOELASTIC; GRANULAR TEMPERATURE: GAS FLUIDIZED BED; CONVECTION CELLS, EXPERIMENTS, FLOW, ONSET, VELOCITY PROFILES: TYPES OF RHEOLOGY

The clustering for relating themes and concepts is exceptionally complex. The categorization taxonomies, and subsequent allocations of phrases among the categories, are functions of the agglomoration technique, association metrics, phrase extraction algorithm, and interpretation of the results. In the present study, the highest level taxonomy was essentially invariant among these parameters, and was used for the examples. Interestingly, it was not substantially different from the highest level taxonomy obtained by visual inspection of the highest frequency phrases, as reported earlier in this paper. To obtain maximum benefits from what clustering can offer, lower categorical hierarchical levels must be accessed. More research is necessary to determine the most desirable combination of parameters to produce clusters at the lower hierarchical levels.

VI. SUMMARY AND CONCLUSIONS

The first two objectives of this study were to demonstrate the feasibility of tracking the myriad impacts of research on other research, development, and applications, using the technical literature, and demonstrate the feasibility of identifying a broad range of research product user characteristics, using the technical literature. Both of these objectives were accomplished, along with some interesting technical insights about vibrating sandpile dynamics and temporal characteristics of information diffusion from research to applications. This wide range of results leads to the following conclusions.

Exploitation of the other types of information contained in the SCI and associated with the citation process offers the potential for providing R&D sponsors information that can help guide future directions of their R&D. In addition, the complete Citation Mining process described in the present paper has the potential to objectively document the breadth of impact of basic research on the R&D community. The addition of text mining to citation bibliometrics will make feasible the large-scale multi-generation citation studies that are necessary to display the full impacts of research.

Text mining is a requirement for making the total Citation Mining possible. Without text mining, either an overly general automated technique, such as journal classification, must be used to identify application areas, or tens or hundreds of thousands of Abstracts must be read. Text mining can locate small numbers of extra-discipline phrases (small signals) from large numbers of intra-discipline phrases (large clutter), and allow only those Abstracts of specific interest to be selected and read.

A substantial amount of human judgement and labor is required for all aspects of Citation Mining. For the bibliometric component of citation mining reported in detail in (20), classifying the results in groupings where judgement is required (e.g., Abstract technical theme, or applications theme) necessitates substantial work. For the text mining component described in detail in this paper, thousands of technical phrases must be examined. Judgements must be made as to their alignment with the main themes of the cited paper(s). Some of the bibliometric components conceivably could be automated (e.g., all the SCI journals could be classified by technical theme beforehand, then the alignment of the cited journal theme to the citing journal theme could be generated automatically). It is not clear how the selection of extra-discipline phrases could be automated, given the intense expert judgement required.

The third of the study objectives was to relate thematic characteristics of citing papers to their cited papers. There was a strong relation of these thematic characteristics for the sandpile paper and its citing papers reported here, and an even stronger citing/ cited paper relationship for the applied photo-voltaic research papers reported in detail in reference (20). This result has potential far-reaching implications for the corporate and national security intelligence communities. Through the tracking of cited papers, one could theoretically infer the theme(s) of the citing papers, and vice versa. Very little has been reported in the literature on this broader field of trans-citation analysis, especially using text mining as reported in the present paper, and the broader field is ripe for further research and exploitation.

This study referred to, but did not examine details of, second or higher generation citations. The authors believe they are valid measures or indicators of influence and impact, but the actual method of impact quantification remains an open question. More research is required to understand the principles of allocating impact among a paper's references.

Finally, there is a very important message that emerges from the results of the present study relative to the sponsorship of basic research. Over the past decade, the trend in industry and government has been toward requirements-driven research (e.g., the term 'strategic research' is becoming used more widely in government agencies, and corporately-funded industrial research has strongly evolved into profit-center sponsored research). While this may be beneficial to the sponsoring organization from a short-term tactical perspective, the long-term strategic perspective may suffer. Would fundamental sand-pile research receive funding from Tokamak, air traffic control, or materials programs, even though sand-pile research could impact these or many other types of applications, as shown in this paper? It is necessary to stress that sponsorship of some unfettered research must be protected, for the strategic long-term benefits on global technology and applications!

VII. REFERENCES FOR SECTION 9

Averch, H., (1994) "Economic Approaches to the Evaluation of Research", Evaluation Review, 18:1, February, p. 77-88.

Braun, T., et al., (1987), "Literature of Analytical Chemistry. A Scientometric Evaluation", CRC Press.

Davidse, R. J., and Van Raan, A. F. J., (1997). “Out of Particles: Impact of CERN, DESY, and SLAC Research to Fields other than Physics”, Scientometrics, 40:2. P. 171-193.

De Solla Price, D. J., (1986) “Little Science, Big Science and Beyond”, Columbia University Press.

Del Río, J. A., Kostoff, R. N., García, E. O., Ramírez, A. M., and Humenik, J. A., (2000). “Citation Mining: Citing Population Profiling using Bibliometrics and Text Mining”. Centro de Investigación en Energía, Universidad Nacional Autonoma de Mexico. http://www.cie.unam.mx/W_Reportes.

DOD, (1969) Project Hindsight, Office of the Director of Defense Research and Engineering, Wash., D. C., DTIC No. AD495905, October.

DOE, (1983) "Health and Environmental Research: Summary of Accomplishments", Office of Energy Research, Office of Program Analysis, Report No. DOE/ER-0194, May; DOE, (1986) "Health and Environmental Research: Summary of Accomplishments", Office of Energy Research, Office of Program Analysis, Report No. DOE/ER-0275, August.

Egghe, L., and Rousseau, R., (1990) “Introduction to Informetrics”, Elsevier.

Herring, S. D., (1999). “The Value of Inter-disciplinarity: A Study Based on the Design of Internet Search Engines”, JASIS, 1 April, p. 358-365.

IDA, (1991) "DARPA Technical Accomplishments", Volume I, IDA Paper P-2192, February 1990; Volume II, IDA Paper P-2429, April 1991; Volume III, IDA Paper P-2538, July 1991, Institute for Defense Analysis.

IITRI, (1968) "Technology in Retrospect and Critical Events in Science", Illinois Institute of Technology Research Institute Report, December.

Jaeger, H. M., and Nagel, S. R. (1992). “Physics of the Granular State”. Science. 256. 20 March, p. 1523-1531.

Kostoff, R. N., (1994) "Assessing Research Impact: US Government Retrospective and Quantitative Approaches", Science and Public Policy, 21:1, February.

Kostoff, R. N., (1997) “The Handbook of Research Impact Assessment”, DTIC Report Number ADA296021, Summer.

Kostoff, R. N., Braun, T., Schubert, A., Toothman, D. R., and Humenik, J. (2000). "Fullerene Roadmaps Using Bibliometrics and Database Tomography". Journal of Chemical Information and Computer Science. 40:1. Jan-Feb. p. 19-39.

Kostoff, R. N., Green, K. A., Toothman, D. R. Humenik, J. A., (2000) “Database Tomography Applied to an Aircraft Science and Technology investment Strategy”, Journal of Aircraft, 37:4, p. 727-730.

Losiewicz, P., Oard, D., and Kostoff, R. N. (2000). "Textual Data Mining to Support Science and Technology Management". Journal of Intelligent Information Systems. 15:2, p. 99-119.

Narin, F., (1989) "The Impact of Different Modes of Research Funding", in: Evered, David and Harnett, Sara, Eds., The Evaluation of Scientific Research, John Wiley and Sons, Chichester, UK, p. 120-140.

Narin, F., Olivastro, D., and Stevens, K. A., (1994) "Bibliometrics -Theory, Practice, and Problems", Evaluation Review, 18:1, February, p. 65-76.

Steele, T. W., (2000). “The Impact of Interdisciplinary Research in the Environmental Science: A Forestry Case Study”, JASIS, 15 March, p. 476-484.

Tassey, G., (1999) “Lessons Learned About the Methodology of Economic Impact Studies: The NIST Experience,” Evaluation and Program Planning, **22, p.** 113—119.

Section 10. Citation Analysis for Assessing Research Performer Quality.

(based on Kostoff, R. N. "Citation Analysis for Research Performer Quality". Scientometrics. 53:1. 49-71. 2002.)

I. OVERVIEW

BACKGROUND: Citation analysis for evaluative purposes typically requires normalization against some control group of similar papers. Selection of this control group is an open question.
OBJECTIVES: Gain a better understanding of control group requirements for credible normalization.
APPROACH: Performed citation analysis on prior publications of two proposing research units, to help estimate team research quality. Compared citations of each unit's publications to citations received by thematically and temporally similar papers.
RESULTS: Identification of thematically similar papers was very complex and labor intensive, even with relatively few control papers selected.
CONCLUSIONS: A credible citation analysis for determining performer or team quality should have the following components:
*Multiple technical experts to average out individual bias and subjectivity
*A process for comparing performer or team output papers with a normalization base of similar papers
*A process for retrieving a substantial fraction of candidate normalization base papers
*Manual evaluation of many candidate normalization base papers to obtain high thematic similarity and statistical representation

II. INTRODUCTION

In the evaluation of science and technology (S&T), whether ongoing or proposed programs, a key criterion is the track record of the proposer or performer. Past analyses [DOE, 1982; Kostoff, 1997a] have shown that, typically, the criterion of Team Quality is the major determinant of program or project quality. Many qualitative and quantitative approaches have been used for the purpose of determining Team Quality [Kostoff, 1997a]. None are viewed as adequate in a stand-alone mode, and present practice is to use multiple approaches to determine Team Quality [Martin, 1983; Kostoff, 1997b].

One of the more widely used of these approaches, especially applicable to research, is citation analysis. For proposer quality assessment, citation analysis consists of counting citations to documents produced by the proposer's research unit, then comparing this citation count to numbers of citations received by similar documents from other research units. The assumption is then made that documents with higher relative numbers of citation counts have more impact than those with lower citation counts, and are of higher quality from the citation metric perspective.

While this approach appears rather straight-forward and deceptively simple, it is intrinsically very complex. This section will illuminate the complexities, and show that high quality S&T citation analysis requires technical experts performing very manually intensive comparisons with very subjective judgements. It will show further that the automated assembly-line approaches to citation analysis, widely used by the decision aid community today, are highly uncertain at low-to-mid citation levels characteristic of most research.

After a background description of the problem, the analytical techniques developed for the citation analysis will be presented. Two illustrative examples of the use of citation analysis to support proposal review will be presented. Because of the confidentiality agreements operable for proposal review, all information that identifies either the proposing organization or the potential science and technology sponsor will be removed. The results of the analysis will then be presented, followed by summary and conclusions that emphasize the lessons learned from using these techniques. Special emphasis will be placed on requirements for thematic similarity between the target documents and the external documents against which they are compared.

III. BACKGROUND

In the present context, citation is referencing, in a document, the work of another individual or group. The work referenced can exist in many forms, although the most common use is reference of another document. Citation analysis is the examination of the multiple dimensions and myriad facets of citations for the purpose of understanding the many impacts of the target documents of interest.

Citation counts resulting from citation analyses are usually classified as outputs, but they are neither outputs nor outcomes. While they are closer to outputs than outcomes, since they can be used in relatively short range analyses and they do not impact the larger problems characteristic of outcomes, they are not under the direct control of the performer.

Modern day interest in studying and developing the citation process accelerated after WW2 [e.g., Zachlin, 1948, Zirkle, 1954]. However, the origins of citation analysis as a widespread bibliometrics tool can be traced to the mid-1950s, with Garfield's proposal for creating a citation index [Garfield, 1955]. As the Science Citation Index (SCI) was developed, along with companion citation indices, the computer revolution and associated information technology developed in parallel. The combination of SCI, massive information storage, and rapid information retrieval laid the foundation for a multi-application S&T evaluation capability.

The foundations of modern traditional citation analysis were established by Garfield [1955, 1963, 1964, 1965, 1966, 1970] and CHI, Inc [Narin, 1975, 1976, 1984, 1994, 1996; Albert, 1991], and extended to co-citation analysis by Small [1973, 1974, 1977, 1981, 1985], Sullivan [1977, 1979, 1980], and Marshakova [1973, 1981, 1988].. The practice of citation analysis has been extended further by groups at the Hungarian Library of Sciences [Schubert, 1986, 1993, 1996; Zsindely, 1982] and the University at Leiden [Moed, 1986; Nederhof, 1987; Braam, 1988, 1991; VanRaan, 1991, 1993, 1996; Davidse, 1997]. A broad summary of the status of citation analysis is contained in a recent festschrift to Eugene Garfield [Festschrift, 2000].

Traditional citation analysis is presently used both at the micro and macro scales. It is used at the micro level, especially in academia, to evaluate components of impact of a given published document, or the documents published by a given researcher or research group. It is used at the macro level to evaluate technical discipline or national outputs. Because of the large numbers of documents and subsequent citations that exist in macro level analyses, semi-automated techniques have been developed to handle the data efficiently. As time has proceeded, these semi-automated techniques have diffused toward micro level application.

Citation analysis has two components. The first component is <u>counting</u> of citations to a document or group of documents, depending on the purpose of the analysis. The second component is placing these citation counts in a

larger context through a comparison and normalization process, to provide meaning to the numbers of counts obtained.

Many articles have been written about problems inherent in the traditional citation analysis process [e.g., Geisler, 2000; MacRoberts, 1989, 1996; Kostoff, 1998]. There are two main categories of problems: those associated with the counts of citations, and those associated with the comparisons of counts of citations. The problems associated with counts of citations can be sub-divided further into problems associated with the quantity of the underlying data, and problems associated with the quality of the underlying data.

III-A. Problems with Citation Counts

III-A-1. Problems with Quantity of Underlying Data

The main resource available for performing citation analysis today is the SCI. The number of candidate articles to be used in a citation analysis is limited to the number of articles in the total SCI. This total is limited by the following sequence of steps.

a) There is approximately $500 billion-$800 billion/ year worth of S&T being performed globally today, depending on one's definition of S&T. Only a small fraction of the S&T performed is documented. While there are many reasons for this [Kostoff, 2000a], basically there are more disincentives to publishing than incentives.

b) Of the S&T performed that eventually gets documented, only a very modest fraction is accessed by the SCI (or any single database). There are tens of thousands each of internal and external technical reports, classified reports and papers, workshop and conference proceedings, journals, magazines, newspapers, and patents resulting from the S&T performed and published annually. Yet, the SCI accesses only about 5600 journals presently. While these accessed journals tend to be the highest quality peer-reviewed research journals, they represent only a fraction of S&T that is documented.

c) Of the documented S&T that is accessed by the SCI, only a fraction reaches the average analyst performing citation analysis. The main reason is

the extremely poor information retrieval techniques actually used by the technical community [Kostoff, 2000b].

Thus, the citation counts derived from the records in the SCI under-represent the total referencing of prior work by the global technical community, and there is no evidence that this under-representation is homogeneous across disciplines or sub-disciplines.

III-A-2. Problems with Quality of Underlying Data

The problems with citation data quality translate into problems with the citation selection process (i.e., the approach used by authors to select references for inclusion in their papers). The issues related to the sociological and cultural aspects of how people cite have been raised by the references cited above, and will not be repeated here. Suffice it to say that the combination of quantity and quality problems with citations places strong limits on the degree to which citations can be used as a stand-alone metric. This is especially true for documents that receive mid and low level numbers of citations (i.e., the vast majority of documents published); the very highly cited documents (a very small fraction of all articles published) are in a class by themselves, and modest margins of error in interpreting their citation counts don't affect overall conclusions about their impact.

III-B. Problems with Citation Comparisons

Problems with citation count comparisons form the focus of this paper. Whether applied to micro or macro scale problems, citation count comparisons have received insufficient attention, and offer further severe constraints on the credibility of present day citation analyses. There are two main types of potential citation count comparisons: comparison of counts to an absolute standard, and comparison of counts to a relative standard. The former comparison is analogous, in the physical sciences, to comparing actual engine efficiencies to maximum engine efficiencies possible (Carnot efficiencies). The latter comparison is analogous to an athletic competition, where one group's performance is compared to another group's performance. One problem with the latter comparison is that the performance of a group is never related to its potential, only to the performance of another 'similar' group. The latter comparison is used in essentially all citation analyses today. This issue of comparison with

absolute or relative standards was examined in a 1997 paper [Kostoff, 1997c], and will not be addressed further.

Citation count comparisons are necessary because of the high variability of citation counts with different parameters. Citation counts depend strongly on the specific technical discipline, or sub-discipline, being examined. The funding and number of active researchers can vary strongly by sub-discipline, and these numbers of researchers affect the numbers of citations directly. The maturity of the sub-discipline affects the numbers of citations, since the basic research community is oriented more toward publishing than the applied research or technology development communities. The breadth of the sub-discipline can affect citation counts, since more focused disciplines will concentrate citations into fewer key researchers. The classification and proprietary levels can vary sharply by sub-discipline, and can strongly affect what gets published and therefore cited in open-literature publications. The documentation and citation culture can vary strongly by sub-discipline. Since citation counts can vary sharply across sub-disciplines, absolute counts have little meaning, especially in the absence of absolute citation count performance standards.

Thus, in order to provide meaning and context to citation counts for performance evaluation in traditional citation analysis, some type of citation count normalization is required. The main normalization approaches used in traditional citation analyses are described in an excellent review article [Schubert, 1996]. They can be summarized as follows:

1) Reference standards based on prior sub-field classification

Journals are classified into a number of science sub-fields. Since some journals are single discipline, and some multi-discipline, percentage weights are assigned to each journal indicating their connection with the different sub-fields. According to Schubert [1996], the method works only at a higher (macro) statistical level; i.e., if the sample under study is large and mixed enough to support the validity of such a statistical approach. Further according to Schubert [1996], for micro level analyses, it is sometimes unavoidable to use a classification scheme concerning not only the journals but every single paper. Schubert proceeds to point out that such classification schemes are enclosed in some specialized databases, such as in the *Physics Briefs*, to classify each paper into one or more of ten first-level and many lower-level sub-fields of physics.

2) Journals as reference standards

Primary journals in science are generally agreed to contain coherent sets of papers both in topics and professional standards. According to Schubert [1996], it seems justified to regard the set of regular authors of a journal as reference standard for any single author (or team of authors), the set of institutions regularly publishing in the journals as reference standard of any single institution, the citation rate of the set of papers published in the journal (or of a properly selected subset) as reference standard of any single paper. Also according to Schubert [1996], one may thus expect that any difference in productivity, citation rate or other scientometric indicators reflects differences in inherent qualities.

3) Related records as reference standards

Subject matter similarity between two documents is measured by the number of shared references. According to Schubert [1996], bibliographic coupling appears to be one of the most selective and flexible techniques of reference standard selection, but "because of its high requirements in time and effort, its use can be suggested only in micro or meso-level".

It is the present author's contention that none of the above normalization methods are adequate for precise normalization, since they do not provide sufficient resolution for distinguishing among the lower level sub-fields. Inability to distinguish precisely among sub-fields translates, in some cases, to substitution of far different magnitude numbers for the normalization base. The next sub-section will show some of the effort required for more precise normalization comparisons.

IV. ANALYSIS TECHNIQUES AND ISSUES

IV-A. First proposal

The author was recently asked, by a potential sponsor, to evaluate an S&T proposal generated by organization XXXX. While there were a number of criteria that had to be evaluated relative to technical quality and relevance of the proposal to the potential sponsor's mission, one key criterion was the quality of the proposer's research team. It was decided to evaluate team quality through evaluation of the research team's various outputs and

outcomes, using citation analysis and other metrics. This section focuses on the citation analysis component used..

The proposal and accompanying material presented many different types of outputs from XXXX researchers. Assessing the quality and impact of those outputs was complex, especially since they covered more than one research area. The following procedure was used as a first-order estimate of quality/ near-term impact of XXXX's output, and thereby of the research team.

The citations of selected XXXX publications were compared against those of thematically similar non-XXXX publications (a control group of publications), using a pair-wise comparison approach. Specifically, all XXXX publications for 1996 (38 documents), as identified in the Web version of the Science Citation Index (SCI), were compared with thematically similar non-XXXX publications from the SCI.

[1996 was selected as a compromise year. The author wanted to examine recent documents that reflected current management and staff of XXXX, but also wanted to insure that sufficient time had passed since publication such that citations had a reasonable chance to accumulate. Figures 1 and 2, titled Citing Papers Time Distribution, show the yearly and cumulative numbers of citing papers as a function of time, for 1996 and 1993, respectively. For 1996, the citing papers (for all the XXXX papers published in 1996) show a linearly increasing cumulative trend up to and including 2000. For 1993, the citing papers (for all the XXXX papers published in 1993) show more of an S-curve trend. While 1993 shows a leveling off of the citations, and would therefore have been a better year to select from that perspective, it was judged to be too far in the past to be relevant for assessing the quality of present XXXX staff and management. Citations from 1996 should almost be ready to level off, if the 1993 distributions can be extrapolated to 1996, and therefore 1996 was selected.]

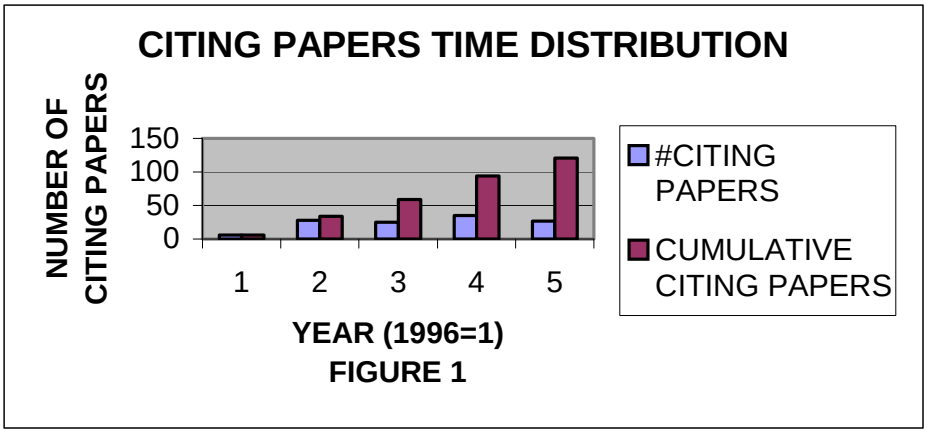


FIGURE 1

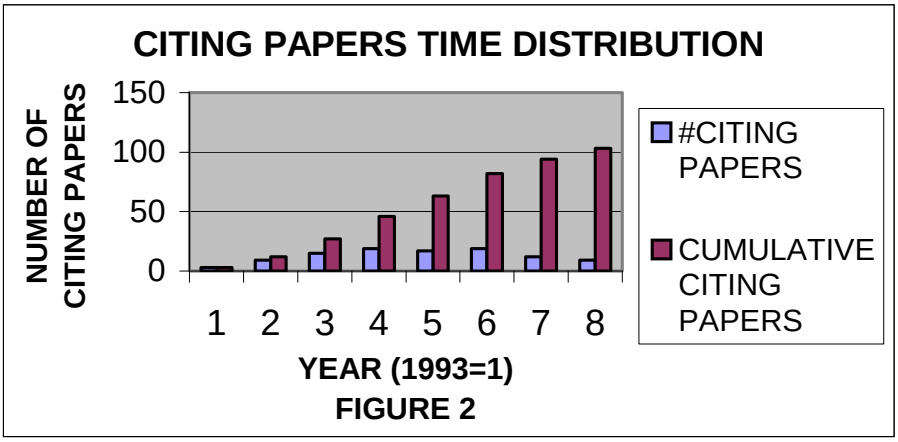


FIGURE 2

Ideally, the size of the control group for each paper should be statistically representative of the total thematically similar non-XXXX papers in the SCI, since the purpose of the citation analysis is to compare the citation performance of each proposer's paper to the aggregate of the relevant performer community.. Practically, resource and time constraints placed

severe linits on the size of the control group. Specifically, for each of the 38 papers published in 1996 (hereafter referred to as the target papers), three non-XXXX papers thematically and temporally similar to the target papers were selected. If 1996 papers with the requisite thematic characteristics could be identified, they were given first priority in the selection, to insure temporal normalization. If 1996 papers could not be identified, then 1997 papers were selected. Thus, the results are conservative with respect to XXXX.

Selection of papers in the SCI thematically similar to the target paper depends strongly on the study's purpose and objectives, the mission of the performing organization, the degree of focus of the paper's theme, the size of the research paper pool from which to choose, and the level of technical description in the paper's SCI Abstract. The relation to study purpose is especially important, and is often overlooked. Specifically, is the purpose of the study to evaluate the 'job right' quality of the performer (i.e., is the specific task selected being performed with the latest tools and techniques to achieve the specific objectives?), or is the purpose of the study to evaluate the 'right job' quality of the performer (i.e., have the right task and right objectives been selected?). If the focus is on 'job right' quality, then the thematically similar papers will be limited to a very narrow area of inquiry. If the focus is on 'right job' quality, then the focus of thematically related papers can be expanded greatly.

For example, suppose that a researcher being evaluated was performing acoustic studies in the 100 KHZ small object detection regime. If the performing organization's mission in acoustics was limited to performing studies only in this regime, and if the quality determination was phrased as how well the researcher was performing relative to other researchers studying the 100 KHZ regime, then the thematically similar papers would all be focused narrowly around frequencies of 100 KHZ. The study reduces to determining the most cited papers at 100 KHZ. If, however, the organization's mission in acoustics provided flexibility in selecting the frequency regime to study, and the organization ***chose*** to focus on the 100 KHZ regime, then thematically related papers could include those in a broader range of frequency regimes. The study reduces to determining the most cited paper in mid-high frequency acoustics. The choice of journal as reference standard, described previously and referenced in Schubert [1996], relates strongly to the latter definition of organization mission, where essentially any paper in an acoustics specialty journal could serve as a

reference standard. The practical implications of 'job right' vs 'right job' comparisons are that papers with substantially higher citation counts could be included in the normalization pool as the allowed definition of thematic similarity becomes broadened.

Selection of papers thematically similar to the target paper was very difficult, time-consuming, and subjective. This was especially true for the broad-based analyses. The selection was more straightforward for the much more limited specific technology papers, since these more focused areas seemed to have many researchers working related problems. The author believes that the subjectivity involved in selecting thematically similar papers is a major source of uncertainty of the results. A rigorous study, in addition to having the rigorous information retrieval and statistical sampling processes mentioned in the next two paragraphs, requires the use of multiple evaluators for the same target papers to average out evaluator subjective bias.

Many of the applied research papers combined analytical technique advancement with novel application advancement. It was not always possible to have thematic similarity for both technique and application, especially in those research areas with relatively few performers, and typically a choice had to be made between technique and application for determining thematic similarity.

Two important issues were i) determining the number of thematically similar candidate papers in the pool from which to choose, and then ii) determining the number of papers to select from the pool. First, in a rigorous study, candidate thematically similar papers would be identified by the most rigorous processes available. In the author's information retrieval studies [Kostoff, 1997d, 2000b], a manually intensive iterative approach using computational linguistics and bibliometrics is used to identify the full scope of relevant literature papers for each specific topic studied. For the present study, this would have required 38 such literature searches. In the time available, even one such rigorous literature search was not feasible. A very approximate approach was used.

Second, the number of papers to select from the candidate pool should have the greatest thematic similarity, and be representative statistically. Again, this would have required poring over hundreds, or thousands, of similar papers, and selecting a substantial number of the most representative

thematically. Again, a small sampling approach was used because of time exigencies.

The first selection step was to examine the Related Records field of the SCI for a given target paper. This field contains papers that have at least one reference in common with the target paper, as stated previously [Schubert, 1996]. Papers that share references tend to be similar thematically, but this is not always true, and the relation between thematic similarity and number of shared references is not always monotonic.

Because of time constraints, a limited number (three) of thematically related papers was examined for each target paper. If three records thematically similar to the target paper could be identified from the Related Records papers, the selection was completed for that target paper. If three records could not be identified, then key words from the target paper's Abstract/ Title/ Keyword fields were used to search the SCI for related records. This approach was substantially more time consuming than the already time-consuming Related Records approach.

FIGURE 3 - CITATION AND FIGURE OF MERIT DATA

A REC#	B PAP CIT	C SELF CIT	D PAP1 CIT	E PAP2 CIT	F PAP3 CIT	G AVER CIT	H FOM1	I MED CITES	J FOM2	K STD DEV CIT	L FOM3
1	4	1	3	3	23	9.667	0.293	3	0.571	11.55	-0.49
2	2	1	9	7	21	12.33	0.14	9	0.182	7.572	-1.36
3	0										
4	0		5	1	2	2.667	0	2	0	2.082	-1.28
5	0		5	6	9	6.667	0	6	0	2.082	-3.2
6	3	2	3	4	4	3.667	0.45	4	0.429	0.577	-1.15
7	0		11	14	4	9.667	0	11	0	5.132	-1.88
8	1	1	1	3	2	2	0.333	2	0.333	1	-1
9	6	3	3	7	5	5	0.545	5	0.545	2	0.5
10	5	0	2	5	16	7.667	0.395	5	0.5	7.371	-0.36
11	5	3	5	2	14	7	0.417	5	0.5	6.245	-0.32
12	2	2	3	3	2	2.667	0.429	3	0.4	0.577	-1.15
13	1	0	4	4	5	4.333	0.188	4	0.2	0.577	-5.77
14	5	2	6	4	9	6.333	0.441	6	0.455	2.517	-0.53
15	7	4	15	5	12	10.67	0.396	12	0.368	5.132	-0.71
16	5	5	3	7	1	3.667	0.577	3	0.625	3.055	0.436
17	4	4	8	4	6	6	0.4	6	0.4	2	-1
18	9	4	38	2	13	17.67	0.338	13	0.409	18.45	-0.47
19	4	2	3	7	7	5.667	0.414	7	0.364	2.309	-0.72
20	2	1	2	6	8	5.333	0.273	6	0.25	3.055	-1.09

21	0	0	2	5	16	7.667	0	5	0	7.371	-1.04
22	1	1	13	8	9	10	0.091	9	0.1	2.646	-3.4
23	24	20	5	2	7	4.667	0.837	5	0.828	2.517	7.682
24	4	0	4	22	8	11.33	0.261	8	0.333	9.452	-0.78
25	0										
26	0										
27	3	0	11	14	2	9	0.25	11	0.214	6.245	-0.96
28	2	2	3	3	4	3.333	0.375	3	0.4	0.577	-2.31
29	4	4	8	10	6	8	0.333	8	0.333	2	-2
30	2	2	3	3	13	6.333	0.24	3	0.4	5.774	-0.75
31	1	1	2	4	5	3.667	0.214	4	0.2	1.528	-1.75
32	0										
33	6	6	13	26	3	14	0.3	13	0.316	11.53	-0.69
34	0	2	2	4		3	0	3	0	1.414	-2.12
35	3	1	2	5	16	7.667	0.281	5	0.375	7.371	-0.63
36	0		2	7	1	3.333	0	2	0	3.215	-1.04
37	2	1	5	22	4	10.33	0.162	5	0.286	10.12	-0.82
38	4	1	5	3	14	7.333	0.353	5	0.444	5.859	-0.57
SUM	115	74	197	200	252	AVER	0.297		0.324		-0.98

Once thematically similar records were identified, the citations for each of the four records were tabulated. Figures of merit were generated, and the citation performance of each target paper was compared with that of the three thematically related papers. The results are shown in Figure 3. Starting from the left, column A is the number of the record, column B is the citations of the target paper, column C is the self-citations of the target paper, columns D, E, F are the citations of the thematically similar papers (the Abstracts of papers 3, 25, 26, 32 did not contain sufficient information for similar papers to be identified), column G is the average citations of the thematically similar papers, column I is the median citations of the thematically similar papers, and column K is the standard deviation of the citations of the thematically similar papers. Columns H, J, L are figures of merit FOM1, FOM2, FOM3, respectively, defined as follows:

FOM1=citations of target paper/ (citations of target paper plus average citations of related papers)

FOM2=citations of target paper/ (citations of target paper plus median citations of related papers)

FOM3=(citations of target paper minus average citations of related papers)/ standard deviations of related papers.

FOM1 and FOM2 have the desirable properties of ranging between zero and unity, as well as equaling 0.5 when the target paper citations equal those of the average or median citations of the related papers. FOM3 removes the limitations of using absolute number values, and places the citation differences in the context of standard deviations.

This section ends with a note about the four papers that could not be evaluated due to insufficient information contained within the Abstract. Ideally, with unlimited time and resources, the full text target and control group papers would be read in their entirety. Practically, time is available for reading Abstracts only. Unfortunately, in the non-medical technical literature, and some of the medical literature, there are no requirements on the technical content of Abstracts. Consequently, many Abstracts contain very little technical detail, and they cannot be used in the citation process. This issue is addressed summarily in a letter to Science [Kostoff, 2001a], and in more detail in a letter to selected technical journal editors proposing the use of Structured Abstracts in all technical journals [Kostoff, 2001b].

IV-B. Second Proposal

In early 1998, the author was asked to evaluate an S&T proposal for a different potential sponsor, generated by an organization (ZZZZ) different from the proposing organization (XXXX) of the first proposal. One critical component again was evaluation of team quality. This was a complex procedure for the second proposal, since most of the organization's publication outputs were co-authored with people from other organizations, and the author wanted to identify the quality of the contributions of researchers from organization ZZZZ only. Again, citation analysis was one of several methods used to gauge team quality, and this section reports on the citation analysis component only.

1. Database Examined and Process Used

One purpose of the study was to examine the citation impact on the technical community of the ZZZZ researchers who publish. Another purpose was to assess some estimate of the ZZZZ researchers' contribution to the published product. Two studies were performed. First, all the 1997 papers in the web version of the SCI that contained a ZZZZ author address were examined. The position of the ZZZZ author in the author list for each paper was highlighted. Citations for this group of papers were not examined, because of the recent date.

Second, all the 1993 papers that contained a ZZZZ author address were examined. 1993 was selected for two reasons. A four-year lag allows many (not all) citations to accumulate, and is sufficient to show differentiation in citation counts among papers. Also, 1993 was the third year that paper abstracts were included in the SCI, allowing more than title information to be obtained about a paper if necessary. Author position was highlighted again, and then the citations received by each paper with citations received by a non-ZZZZ authored paper of similar theme were compared.

V. RESULTS AND DISCUSSION

V-A. First Proposal

The results for the first proposal are as follows.

Figures 4 and 5, titled Citation Distribution Function, show the numbers of papers N(X) with X cites for 1993 and 1996, respectively. 63% of the 1993 target papers had either zero or one cites, and 37% of the 1996 target papers had either zero or one cites. For 1996, the average number of citations per target paper was three, of which 2/3 were self-cites. (No judgements are made about including or excluding self-cites. To make such judgements rationally, each full-text paper would have to be read, and the technical rationale for self-citation other than author self-gratification would have to made. Such a level of detail is beyond the scope of this study.) For 1993, the average number of citations per target paper was about 2.5. For 1996, the average number of citations per thematically related paper was about twice the number of target paper citations.

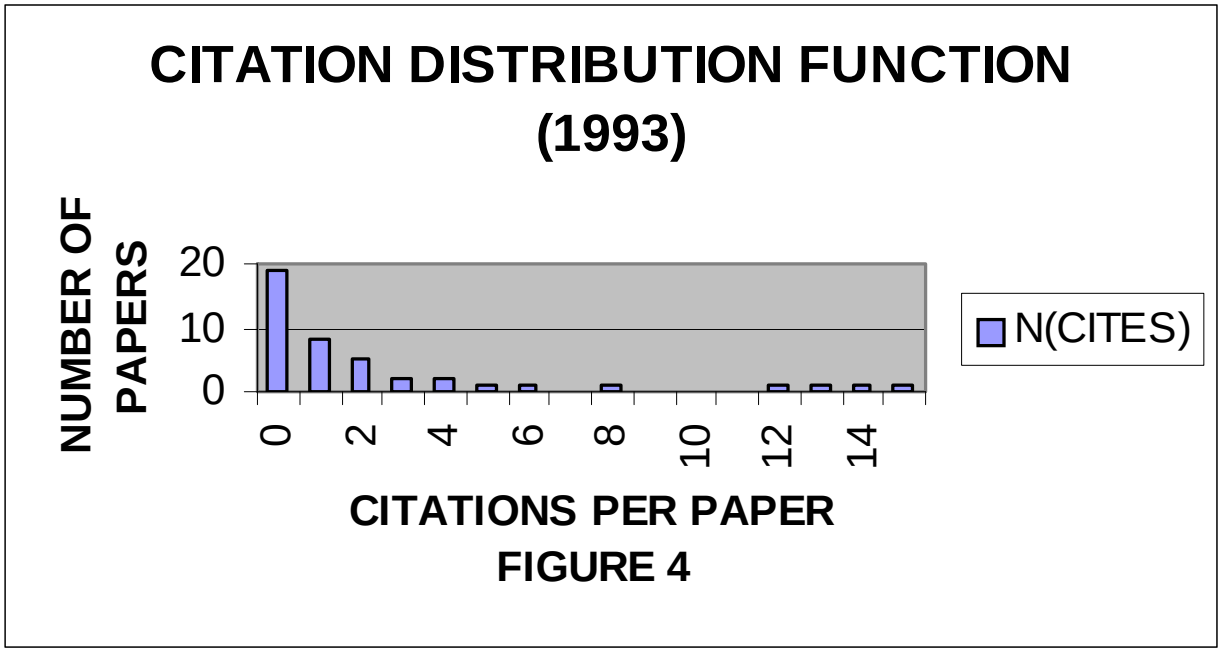

CITATION DISTRIBUTION FUNCTION
(1993)
NUMBER OF PAPERS
20
10
0
0
2
4
6
8
10
12
14
CITATIONS PER PAPER
N(CITES)

FIGURE 4

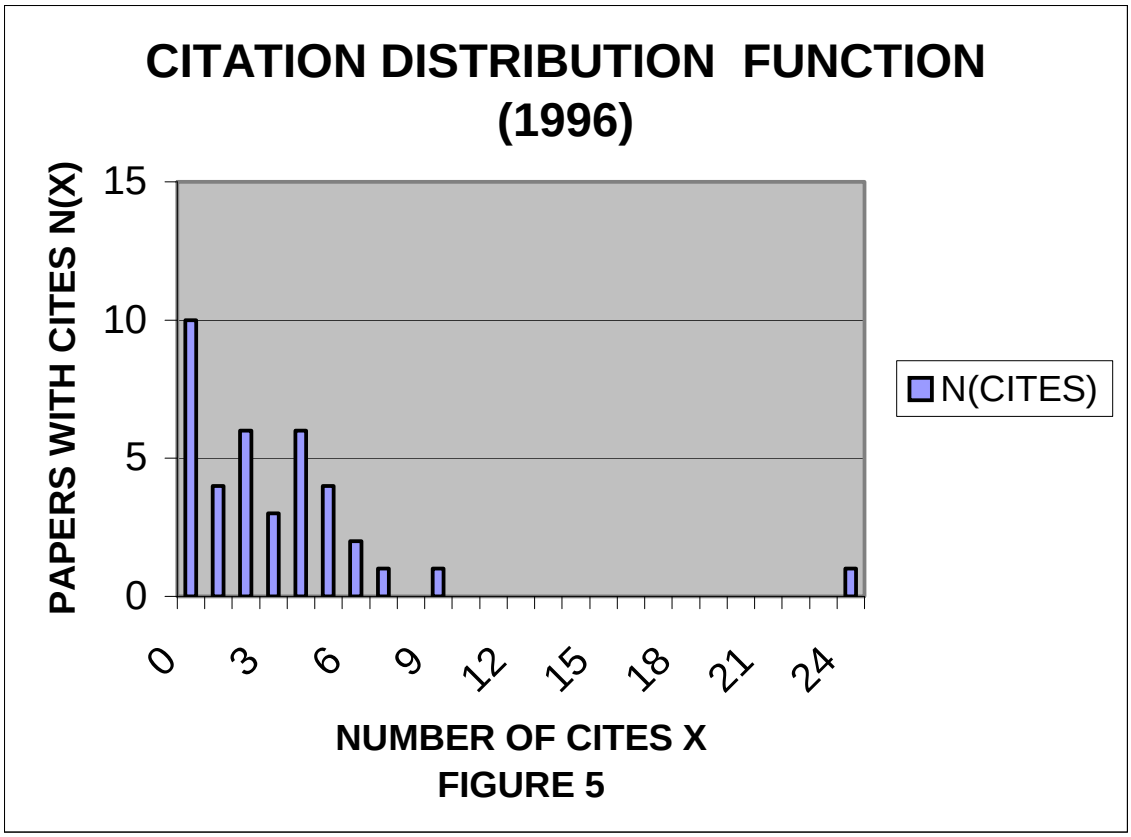


FIGURE 5

For 1996, the average value of FOM1 and FOM2 was about 0.3, and the average value of FOM3 was about minus one standard deviation. Thus, all three figures of merit gave essentially similar results. FOM1 and FOM2 were greater than 0.5 in less than ten percent of the target papers examined. In the best performing target paper, both in absolute citations and relative citations, 20 of the 24 citations were self-cites. This particular paper had many authors, and many of these authors cited the target paper in later publications.

Many of the research disciplines examined seem to have relatively few papers thematically related to the target paper. In addition, the absolute levels of citations are low, relative to other disciplines the author has examined. This suggests research into areas that have few performers, probably low funding, and therefore low citations.

V-B. Second Proposal

1. Results and Discussion

a. 1997 Database

In the 1997 database, there were 43 papers in the SCI with a ZZZZ address for the research unit. These papers had a total of 184 authors, with an average of 4.29 authors per paper, a median of 3 authors per paper, and a mode of 3 authors per paper. A Coefficient of Author Position (CAP) was defined as a measure of the ZZZZ author's location in the total author list. The definition of CAP was:

CAP=(x-1)/(n-1)

where x was the location of the ZZZZ author in the list, and n was the total number of authors in the list. Thus, if there were three authors in the list, and the ZZZZ author was third, CAP would equal one. If the ZZZZ author was first in this case, CAP would equal zero. If the paper had only one author, CAP was set equal to zero. Thus, the higher the value of CAP, the less was the relative contribution of the ZZZZ author.

There are two assumptions here. First, the ordinal positioning of any author in the list reflects his/ her relative contribution to the paper. In the absence of large power differential relationships (e.g., advisor/ student), this is probably a very reasonable assumption. In the presence of large power differential relationships, it may or may not be reasonable, but validation of the assumption would be next to impossible.

Second, the ordinal positioning can be quantified for computational purposes. There appears to be nothing in the literature that supports or rejects this assumption. For large numbers of papers undergoing citation analyses, anomolies will disappear, and quantification for estimation purposes may be reasonable. However, because of the uncertainty of the validity of this assumption, supplementary approaches were used to estimate the contribution of organization ZZZZ's researchers to overall paper quality. In this particular case, there were no significant differences in final results among the different methods used.

The total value of CAP summed over the 43 papers was 26.27, with an average value of 0.61, a median value of .92, and a mode of 1. Most papers were multi-authored; there were only four papers with one author. To summarize these results, the preponderance of papers that include an ZZZZ research unit author address have multiple authors, and the ZZZZ author is usually at the

end of this list. The typical paper in this database had about three authors, with the ZZZZ author being last.

b. 1993 Database

i. Author Position Study

In the 1993 database, there were 44 papers in the SCI with an ZZZZ address. These papers had a total of 126 authors, with an average of 2.86 authors per paper, a median of 3 authors per paper, and a mode of 3 authors per paper. The total value of CAP summed over the 44 papers was 18.97, with an average value of .43, a median double value of 0/.5 (half the papers had a CAP of zero, the other half had a CAP of .5 or greater) and a mode of 0. The typical paper in this database had about three authors, with the ZZZZ author being second.

In comparison with the 1997 database results, the total number of papers is about the same. The median and mode of authors per paper is the same, but the average has dropped by a third from 1997 papers to 1993 papers. More importantly, the average CAP value dropped by a third from 1997 to 1993, the median CAP value dropped by a half, and the mode plummeted from one to zero. Thus, in 1993, the ZZZZ authors were contributing significantly more to papers (as measured by their ordinal position in the authors list) than in 1997.

ii. Citation Comparison Study

For the 1993 database, citations of pairs of similar theme papers were compared. In particular, for a given paper with a ZZZZ author address in the list, a similar theme paper was selected from the Related Records field, and the number of citations received by each paper was transcribed and compared. The procedure used was to select the first 1993 paper from the Related Records field with a similar theme to the target paper (this procedure normalized publication date and theme), and compare each paper's citations. (In a very few cases, no 1993 papers could be found in the Related Records field, and a 1994 or 1992 paper of similar theme was used. In a very few cases, no similar theme paper could be found for 1992 or 1994.)

Then, the ratio of citations of the two papers was transcribed, and this ratio was placed in one of five bands: very high (VH), high (H), same (S), low (L), very low (VL).

'Very High', for example, meant that the ratio of citations received by the related paper to the citations received by the ZZZZ paper was very high, a subjective judgement made by observation. 'Same' meant that the numbers of citations received by the two papers were close, not necessarily identical. Typically, citations received by a few of the other related papers would be examined to ascertain the approximate range of citations, and then judgements about the significance of the differences in citation numbers would be made. Obviously, in a definitive or final study of this nature, there would need to be people involved who could judge if in fact themes were closely related, and there would need to be citation distribution studies of related papers to obtain a more quantitative basis for judging significance of differences.

The population of the five bands was as follows: 12(VH); 9(H); 14(S); 4(L); 1(VL), for a total of 40 pairs where the citations could be compared. While the mode is in the S band, the median is in the H band. Since half the papers in the database had a CAP of zero, all other things being equal one would expect six papers in the VH band to have a CAP of zero. In actuality, nine papers in the VH band had a CAP of zero. Thus, those papers with a VH figure of merit tended to have more ZZZZ lead authors than one would expect from the database overall average.

There were seven prolific ZZZZ authors, each of whom participated in three or more papers. The population of the five bands for these seven prolific authors was: 1(VH); 5(H); 9(S); 3(L); 0(VL). Compared to the overall 1993 database, where 52.5% of the ZZZZ papers were in the VH or H bands, these seven authors had 33% of papers in the VH and H bands. Also, for these seven authors, the average CAP was .6, the median CAP was 0.8, and the mode CAP was 1. For the 1993 database, the parallel numbers were .43 (av), 0/.5 (med), 0 (mode). Thus, while the more prolific authors had better relative citeability than the database average, these authors were closer to the end of the author listing than the database average.

iii. Discussion

The highlights of this author position study are:
* The preponderance of 1997 papers that include a ZZZZ author address have multiple authors, and the ZZZZ author is usually at the end of this list. The typical paper in this database had about three authors, with the ZZZZ author being last.

* In 1993, the ZZZZ authors were contributing significantly more to papers (as measured by their ordinal position in the authors list) than in 1997. The typical paper in the 1993 database had about three authors, with the ZZZZ author being second.
* Those papers with a VH figure of merit tended to have more ZZZZ lead authors than one would expect from the database overall average.
* While the more prolific ZZZZ authors in 1993 had better relative citeability than the database average, these authors were closer to the end of the author listing than the database average.
* More work needs to be done to place ordinal position quantification on a stronger scientific foundation.

In about half the cases, papers with a ZZZZ author address were cited as well as, or better than, comparable non-ZZZZ address papers. On the surface, it appears that papers with ZZZZ authors are having a reasonable impact on the technical community. However, the contribution of the ZZZZ authors to these papers, especially those where the ZZZZ author is listed last, remains unknown. It would have been useful to compare the number of authors for each paper in the pair; this might have shed some light on whether or not the ZZZZ papers are 'author heavy'. This was not done because this issue was not recognized until now. It would also be useful to ascertain why the ZZZZ authors dropped back in their ordinal position in the author list from 1993 to 1997.

VI. SUMMARY AND CONCLUSIONS

This section has provided two examples of the application of citation analysis to proposal evaluation. A number of lessons were learned concerning requirements for high quality citation analysis. These lessons are summarized as follows.

A. Since citation counts can vary sharply across sub-disciplines, absolute counts have little meaning, especially in the absence of absolute citation count performance standards. In order to provide meaning and context of citation counts for performance evaluation in citation analysis, some type of citation count normalization is required.

B. Three types of reference standards are used traditionally for citation analysis: 1) Reference standards based on prior sub-field classification; 2) Journals as reference standards; 3) Related records as reference standards.

None of the above normalization methods are adequate for precise normalization, since they do not provide sufficient resolution for distinguishing among the lower level sub-fields. Inability to distinguish precisely among sub-fields translates, in some cases, to substitution of far different magnitude numbers for the normalization base

C. Selection of papers in the SCI thematically similar to the target paper depends strongly on the study's purpose and objectives, the mission of the performing organization, the degree of focus of the paper's theme, the size of the research paper pool from which to choose, and the level of technical description in the paper's SCI Abstract. The relation to study purpose is especially important, and is often overlooked. If the focus is on 'job right' quality, then the thematically similar papers will be limited to a very narrow area of inquiry. If the focus is on 'right job' quality, then the focus of thematically related papers can be expanded greatly. The practical implications of 'job right' vs 'right job' comparisons are that papers with substantially higher citation counts could be included in the normalization pool as the allowed definition of thematic similarity becomes broadened.

D. Selection of papers thematically similar to the target paper was very difficult, time-consuming, and subjective. This was especially true for the broad-based analyses. The selection was more straightforward for the much more limited specific technology papers, since these more focused areas seemed to have many researchers working related problems. The subjectivity involved in selecting thematically similar papers is a major source of uncertainty of the results. A rigorous study, in addition to having the rigorous information retrieval and statistical sampling processes mentioned in the next two paragraphs, requires the use of multiple evaluators for the same target papers to average out bias.

E. Many of the applied research target papers combined analytical technique advancement with novel application advancement. It was not always possible to have thematic similarity for both technique and application, especially in those research areas with relatively few performers. Typically, a choice had to be made between technique and application for determining thematic similarity.

F. Two important issues were i) determining the number of thematically similar candidate papers in the pool from which to choose, and then ii) determining the number of papers to select from the pool. First, in a credible

study, candidate thematically similar papers would be identified by the most rigorous processes available, and such processes are presently very complex and time-consuming. Second, the number of papers to select from the candidate pool should have the greatest thematic similarity, and be representative statistically. Such selection would have required poring over hundreds, or thousands, of similar papers, and selecting a substantial number of the most representative thematically.

G. Contrary to much popular thinking, the technical expertise of the citation analyst can have a major impact on the quality of the results. The type of pair-wise comparison required for credible citation studies is a highly subjective process, requiring the selection of a thematically similar normalization base. If the analyst understands the subject matter, the subjective judgements made will be reasonably accurate. If the analyst is not a technical expert in the subject area, the results will contain a high degree of uncertainty. Thus, in a rigorous citation analysis, multiple technical experts are necessary to average out individual bias and subjectivity, and much manually intensive effort is required for the normalization process.

Operationally, the above results suggest that a credible citation analysis for determining performer or team quality should have the following components:
*Multiple technical experts to average out individual bias and subjectivity
*A process for comparing performer or team output papers with a normalization base of similar papers
*A process for retrieving a substantial fraction of candidate normalization base papers
*Manual evaluation of many candidate normalization base papers to obtain high thematic similarity and statistical representation

Since the use of citation analysis as one metric for determining research performer or team quality is substantially under-utilized in government and industry at present, the addition of the above requirements to the citation analysis process would only serve to reduce its utilization further. Pragmatically, tradeoffs are required if citation analysis is to be used as an evaluative tool. The degradation in citation analysis quality as the above conditions are relaxed needs to be studied further.

VII. REFERENCES FOR SECTION 10

Albert, M.B., Avery, D., Narin F., Mcallister P., “Direct Validation Of Citation Counts As Indicators Of Industrially Important Patents” , Research Policy 20: (3) 251-259 , Jun 1991.

Braam, R.R., Moed H.F., Vanraan A.F.J., “Mapping Of Science By Combined Co-Citation And Word Analysis .1. Structural Aspects” , Science Technology & Human Values 13: (1-2) 97-98 ,Win-Spr 1988.

Braam, R.R., Moed, H.F., Vanraan A.F.J.,“Mapping Of Science By Combined Cocitation And Word Analysis .1. Structural Aspects”, Journal Of The American Society For Information Science 42: (4) 233-251, May 1991.

Davidse, R. J., and VanRaan, A. F. J., “Out of Particles: Impact of CERN, DESY, and SLAC Research to Fields other than Physics”, Scientometrics, 40:2. P. 171-193, 1997.

Del Río, J. A., Kostoff, R. N., García, E. O., Ramírez, A. M., and Humenik, J. A., “Citation Mining Citing Population Profiling using Bibliometrics and Text Mining”. Centro de Investigación en Energía, Universidad Nacional Autonoma de Mexico, 2001. http://www.cie.unam.mx/W_Reportes.

DOE, "An Assessment of the Basic Energy Sciences Program", Office of Energy Research, Office of Program Analysis, Report No. DOE/ER-0123, March 1982.

Festschrift, ASIST Monograph Series, Web Of Knowledge - A Festschrift In Honor Of Eugene Garfield, 2000.

Garfield, E., “Citation Indexes For Science - New Dimension In Documentation Through Association Of Ideas”, Science, 122: (3159) 108-111, 1955.

Garfield , E., Sher , I. H., “New Factors In Evaluation Of Scientific Literature Through Citation Indexing” American Documentation, 14: (3) , 1963.

Garfield , E., “Science Citation Index-New Dimension In Indexing Unique Approach Underlies Versatile Bibliographic Systems For Communicating + Evaluating Information”, Science, 144: (361), 1964.

Garfield , E., “Can Citation Indexing Be Automated”, Statistical Association Methods For Mechanized Documentation Symposium Proceedings 1964: (Nbs26) 189, 1965.

Garfield E., “Patent Citation Indexing And Notions Of Novelty Similarity And Relevance”, Journal Of Chemical Documentation 6: (2), 1966.

Garfield., E., “Citation Indexing For Studying Science”, Nature 227: (5259) , 1970.

Geisler, E., “The Metrics of Science and Technology”, Quorum Books, Westport, CT, 2000.

Kostoff, R. N., "The Handbook of Research Impact Assessment," Seventh Edition, Summer 1997, DTIC Report Number ADA296021. Also, available at http://www.dtic.mil/dtic/kostoff/index.html, 1997a.

Kostoff, R. N., "Peer Review: The Appropriate GPRA Metric for Research", Science, Volume 277, 1 August 1997b.

Kostoff, R. N., "Citation Analysis Cross-Field Normalization: A New Paradigm", Scientometrics, 39:3, 1997c.

Kostoff, R. N., Eberhart, H. J., and Toothman, D. R., "Database Tomography for Information Retrieval", Journal of Information Science, 23:4, 1997d.

Kostoff, R. N., "The Use and Misuse of Citation Analysis in Research Evaluation", Scientometrics, 43:1, September, 1998.

Kostoff, R. N., “The Underpublishing of Science and Technology Results”, The Scientist, 1 May 2000a.

Kostoff, R. N., "High Quality Information Retrieval for Improving the Conduct and Management of Research and Development", Proceedings:

Twelfth International Symposium on Methodologies for Intelligent Systems, 11-14 October 2000b.

Kostoff, R. N., and Hartley, J., “Structured Abstracts For Technical Journals”, Science, 11 May 2001a.

Kostoff, R. N., and Hartley, J., “Structured Abstracts For Technical Journal Articles, Letter to Technical Journal Editors, 14 May 2000b. Letter available from author.

MacRoberts, M.H., and MacRoberts, B.R., "Problems of Citation Analysis: A Critical Review," Journal of the American Society for Information Science, 40:5, 1989.

MacRoberts, M., and MacRoberts, B., "Problems of Citation Analysis", Scientometrics, 36:3, July-August, 1996.

Marshakova.IV , “System Of Document Connections Based On References”, Nauchno-Tekhnicheskaya Informatsiya Seriya 2-Informatsionnye Protsessy I Sistemy, (6) 3-8 1973

Marshakova IV , “Citation Networks In Information-Science”, Scientometrics, 3: (1) 13-25 1981

Marshakova IV , “On The Mapping Of Science”, Vestnik Akademii Nauk Sssr, (5) 70-82 1988

Martin, B. and Irvine, J., "Assessing Basic Research: Some Partial Indicators of Scientific Progress in Radio-Astronomy," Research Policy, 12, 1983.

Moed H.F., Vanraan A.F.J., “Observations And Hypotheses On The Phenomenon Of Multiple Citation To A Research Groups Oeuvre.”, Scientometrics 10: (1-2) 17-33 , July 1986.

Narin. , F., Carpenter M.P., “National Publication And Citation Comparisons”, Journal Of The American Society For Information Science 26: (2) 80-93 , 1975.

Narin, F., "Evaluative Bibliometrics: The Use of Publication and Citation Analysis in the Evaluation of Scientific Activity" (monograph), NSF C-637,

National Science Foundation, Contract NSF C-627, NTIS Accession No. PB252339/AS, March 31, 1976.

Narin, F., Carpenter M.P., Woolf P., “Technological Performance Assessments Based On Patents And Patent Citations”, IEEE Transactions On Engineering Management 31: (4) 172-183 , 1984

Narin, F., Olivastro, D., and Stevens, K. A., "Bibliometrics -Theory, Practice, and Problems", in: Kostoff, R. N., (ed.), Evaluation Review, Special Issue on Research Impact Assessment, 18:1, February 1994.

Narin F., Hamilton K.S., “Bibliometric Performance Measures”, Scientometrics 36: (3) 293-310 , Jul-Aug 1996.

Nederhof A.J., Vanraan A,F.J., “Citation Theory And The Ortega Hypothesis”, Scientometrics 12: (5-6) 325-328 , Nov 1987

Schubert A., Glanzel W. , Braun T., “Relative Indicators Of Publication Output And Citation Impact Of European Physics Research: 1978-1980”, Czechoslovak Journal Of Physics 36: (1) 126-129 , 1986

Schubert A., Braun T., ”Reference-Standards For Citation Based Assessments”, Scientometrics 26: (1) 21-35 , Jan 1993

Schubert, A., and Braun, T., "Cross-Field Normalization of Scientometric Indicators", Scientometrics, 36:3, 1996.

Small , H. G. “Relationship Between Citation Indexing And Word Indexing - Study Of Co-Occurrences Of Title Words And Cited References”, Proceedings Of The American Society For Information Science 10: 217-218 , 1973.

Small , H., “Co-Citation In Scientific Literature - New Measure Of Relationship Between 2 Documents”, Current Contents (7) 7-10, 1974.

Small , H. G., “Co-Citation Model Of A Scientific Specialty - Longitudinal-Study Of Collagen Research”, Social Studies Of Science 7: (2) 139-166, 1977.

Small. , H., “The Relationship Of Information-Science To The Social-

Sciences - A Co-Citation Analysis", Information Processing & Management 17: (1) 39-50, 1981.

Small H., Sweeney E., Greenlee E., "Clustering The Science Citation Index Using Co-Citations .2. Mapping Science", Scientometrics , 8, 1985.

Sullivan D, White Dh, Barboni Ej, "Co-Citation Analyses Of Science – Evaluation", Social Studies Of Science, 7: (2) 223-240 1977

Sullivan D, Koester D, White Dh, Kern R, "Understanding Rapid Theoretical Change In Particle Physics - Month-By-Month Co-Citation Analysis," Proceedings Of The American Society For Information Science, 16: 276-285 1979

Sullivan D, Koester D, White Dh, Kern R, "Understanding Rapid Theoretical Change In Particle Physics - A Month-By-Month Co-Citation Analysis," Scientometrics, 2: (4) 309-319 1980

Vanraan, A.F.J., "Fractal Geometry of Information Space As Represented By Co-Citation-Clustering" , Scientometrics 20: (3) 439-449 , Mar-Apr 1991.

Vanraan, A.F.J., Tijssen R.J.W., "The Neural Net Of Neural Network Research - An Exercise In Bibliometric Mapping" , Scientometrics 26: (1) 169-192 , Jan 1993.

Vanraan, A.F.J., "Advanced Bibliometric Methods As Quantitative Core Of Peer Review Based Evaluation And Foresight Exercises", Scientometrics 36: (3) 397-420 , Jul-Aug, 1996.

Zachlin, A. C., "On Literature Citation", Science, 107: (2777) 292-293, 1948.

Zirkle, C., "Citation of Fraudulent Data", Science, 120: (3109) 189-190, 1954.

Zsindely, S., Schubert A., Braun T., "Citation Patterns Of Editorial Gatekeepers In International Chemistry Jour

Section 11. Citation-Assisted Background

(based on Kostoff, R.N., and Shlesinger, M. F. "CAB-Citation-Assisted Background." Scientometrics. 62:2. 199-212. 2005.)

OVERVIEW

A chronically weak area in research papers, reports, and reviews is the complete identification of background documents that formed the building blocks for these papers. A method for systematically determining these seminal references is presented. Citation-Assisted Background (CAB) is based on the assumption that seminal documents tend to be highly cited. CAB is being applied presently to three applications studies (Anthrax, nanotechnology, high speed flow), and the results so far are much superior to those used by the first author for background development in any other study. An example of the application of CAB to the field of Nonlinear Dynamics is outlined. While CAB is a highly systematic approach for identifying seminal references, it is not a substitute for the judgment of the researchers, and serves as a supplement.

INTRODUCTION

Research is a method of systematically exploring the unknown to acquire knowledge and understanding. Efficient research requires awareness of all prior research and technology that could impact the research topic of interest, and builds upon these past advances to create discovery and new advances. The importance of this awareness of prior art is recognized throughout the research community. It is expressed in diverse ways, including requirements for Background sections in journal research articles, invited literature surveys in targeted research areas, and required descriptions of prior art in patent applications.

For the most part, development of Background material for any of the above applications is relatively slow and labor intensive, and limited in scope. Background material development usually involves some combination of manually sifting through outputs of massive computer searches, manually tracking references through multiple generations, and searching ones own records for personal references. The few studies that have been done on the adequacy of Background material in documents show that only a modest fraction of relevant material is included (MacRoberts and MacRoberts, 1989,

1996; Liu, 1993; Calne and Calne, 1992; Shadish et al, 1995; Moravcsik and Murugesan, 1975).

In particular, an analysis of Medline papers on the haemodynamic response to orotracheal intubation showed that recognized deficiencies in research method were not acknowledged. The authors recommended that, when submitting work for publication, investigators should provide evidence of how they searched for previous work (Smith and Goodman, 1997).

Another specific example was provided by MacRoberts and MacRoberts (1997). Replicating their earlier work in a journal on genetics which indicated that only 30% of influences evident in text are reflected in a paper's references, the text of an issue of Sida was studied by the MacRoberts to extract influences of previous work evident therein. Influences they judged present in the text appeared in the references only 29% of the time.

Typically missing from standard Background section or review article development, as well as in the specific examples cited above, is a systematic approach for identifying the key documents and events that provided the groundwork for the research topic of interest. The present section presents such a systematic approach for identifying the key documents, called Citation-Assisted Background (CAB). The next sub-section describes the CAB concept, and provides an outline of its operation, with an illustrative example from the research area of Nonlinear Dynamics.

CONCEPT DESCRIPTION

The CAB concept identifies the key Background documents for a research area using citation analysis. CAB rests on the assumption that a document that is a significant building block for a specific research area will typically have been referenced positively by a substantial number of people who are active researchers in that specific area. Implementation of the CAB concept then requires the following steps:

- The research area of interest must be defined clearly
- The documents that define the area of interest must be identified and retrieved

- The references most frequently used in these documents must be identified and selected
- These critical references must be analyzed, and integrated in a cohesive narrative manner to form a comprehensive Background section or separate literature survey

These required steps are achieved in the following manner.

1. The research topic of interest is defined clearly by the researchers who are documenting their study results. For example, consider the research area of Nonlinear Dynamics. In a recent text mining study of Nonlinear Dynamics (Kostoff et al, 2004), the research area was defined as "that class of motions in deterministic physical and mathematical systems whose time evolution has a sensitive dependence on initial conditions."
2. The topical definition is sharpened further by the development of a literature retrieval query. In the text mining study mentioned above, the literature retrieval query was ((CHAO* AND (SYSTEM* OR DYNAMIC* OR PERIODIC* OR NONLINEAR OR BIFURCATION* OR MOTION* OR OSCILLAT* OR CONTROL* OR EQUATION* OR FEEDBACK* OR LYAPUNOV OR MAP* OR ORBIT* OR ALGORITHM* OR HAMILTONIAN OR LIMIT* OR QUANTUM OR REGIME* OR REGION* OR SERIES OR SIMULATION* OR THEORY OR COMMUNICATION* OR COMPLEX* OR CONVECTION OR CORRELATION* OR COUPLING OR CYCLE* OR DETERMINISTIC OR DIMENSION* OR DISTRIBUTION* OR DUFFING OR ENTROPY OR EQUILIBRIUM OR FLUCTUATION* OR FRACTAL* OR INITIAL CONDITION* OR INVARIANT* OR LASER* OR LOGISTIC OR LORENZ OR MAGNETIC FIELD* OR MECHANISM* OR MODES OR NETWORK* OR ONSET OR TIME OR FREQUENC* OR POPULATION* OR STABLE OR ADAPTIVE OR CIRCUIT* OR DISSIPAT* OR EVOLUTION OR EXPERIMENTAL OR GROWTH OR HARMONIC* OR HOMOCLINIC OR INSTABILIT* OR OPTICAL)) OR (BIFURCATION* AND (NONLINEAR OR HOMOCLINIC OR QUASIPERIODIC OR QUASI-PERIODIC OR DOUBLING OR DYNAMICAL SYSTEM* OR EVOLUTION OR INSTABILIT* OR SADDLE-NODE* OR MOTION* OR OSCILLAT* OR TRANSCRITICAL OR BISTABILITY OR LIMIT CYCLE* OR POINCARE OR LYAPUNOV OR ORBIT*)) OR (NONLINEAR AND (PERIODIC SOLUTION* OR OSCILLAT* OR MOTION* OR

HOMOCLINIC)) OR (DYNAMICAL SYSTEM* AND (NONLINEAR OR STOCHASTIC OR NON-LINEAR)) OR ATTRACTOR* OR PERIOD DOUBLING* OR CORRELATION DIMENSION* OR LYAPUNOV EXPONENT* OR PERIODIC ORBIT* OR NONLINEAR DYNAMICAL) NOT (CHAO OR CHAOBOR* OR CHAOTROP* OR CAROTID OR ARTERY OR STENOSIS OR PULMONARY OR VASCULAR OR ANEURYSM* OR ARTERIES OR VEIN* OR TUMOR* OR SURGERY)

3. The query is entered into a database search engine, and documents relevant to the topic are retrieved. In the text mining study mentioned above, 6160 documents were retrieved from the Web version of the Science Citation Index (SCI) for the year 2001. The SCI was used because it is the only major research database to contain references, in a readily extractable format.
4. These documents are combined to create a separate database, and all the references contained in these documents are extracted. Identical references are combined, the number of occurrences of each reference is tabulated, and a table of references and their occurrence frequencies is constructed. In the text mining study on Nonlinear Dynamics, 113176 separate references were extracted and tabulated. Table 1 contains the twenty highest frequency (most cited) references extracted from the Nonlinear Dynamics database.

TABLE 1 – MOST HIGHLY CITED DOCUMENTS

AUTHOR	YEAR	SOURCE	VOL	PAGE	# CIT
PECORA LM	1990	PHYS REV LETT	V64	P821	177
GUCKENHEIMER J	1983	NONLINEAR OSCILLATIO			149
OTT E	1990	PHYS REV LETT	V64	P1196	142
LORENZ EN	1963	J ATMOS SCI	V20	P130	115
CROSS MC	1993	REV MOD PHYS	V65	P851	105
WOLF A	1985	PHYSICA D	V16	P285	103
TAKENS F	1981	LECT NOTES MATH	V898	P366	97
OTT E	1993	CHAOS DYNAMICAL SYST			97
GRASSBERGER P	1983	PHYSICA D	V9	P189	94
GUTZWILLER MC	1990	CHAOS CLASSICAL QUAN			88
ROSENBLUM MG	1996	PHYS REV LETT	V76	P1804	77
GRASSBERGER P	1983	PHYS REV LETT	V50	P346	76
ECKMANN JP	1985	REV MOD PHYS	V57	P617	75
THEILER J	1992	PHYSICA D	V58	P77	66
NAYFEH AH	1979	NONLINEAR OSCILLATIO			62
FUJISAKA H	1983	PROG THEOR PHYS	V69	P32	61
WIGGINS S	1990	INTRO APPL NONLINEAR			61
RULKOV NF	1995	PHYS REV E	V51	P980	59

PYRAGAS K	1992	PHYS LETT A	V170	P421	59
LICHTENBERG AJ	1992	REGULAR CHAOTIC DYNA			58

Two frequencies are computed for each reference, but only the first is shown in Table 1. The frequency shown in the rightmost column is the number of times each reference was cited by the 6160 records in the retrieved database only. This number reflects the importance of a given reference to the specific discipline of Nonlinear Dynamics. The second frequency number (not shown) is the total number of citations the reference received from all sources, and reflects the importance of a given reference to all the fields of science that cited the reference. This number is obtained from the citation field or citation window in the SCI. In CAB, only the first frequency is used, since it is topic-specific. Using the first discipline-specific frequency number obviates the need to normalize citation frequencies for different disciplines (due to different levels of activity in different disciplines), as would be the case if total citation frequencies were used to determine the ordering of the references.

Before presenting a specific implementation algorithm for the Nonlinear Dynamics example, a few caveats will be discussed. First, listing and selection of the most highly cited references are dependent on the comprehensiveness and balance of the total records retrieved. Any imbalances (from skewed databases or incorrect queries) can influence the weightings of particular references, and result in some references exceeding the selection threshold where not warranted, and others falling below the threshold where not warranted.

Second, it is important that the query used for record retrieval be extensive (Khan and Khor, 2004; Harter and Hert, 1997; Kantor, 1994), as was shown for the Nonlinear Dynamics example. The query needs to be checked for precision and recall, which becomes complicated when assumptions of binary relevance and binary retrieval are relaxed (Della Mea and Mizzaro, 2004). There are a multitude of issues to be considered when evaluating queries and their impact on precision and recall. A recent systems analytic approach to analyzing the information retrieval process concludes that, for completeness, the interaction of the Environment and the information retrieval system must be considered in query development (Kagolovsky and Moehr, 2004). The first author's experiences (with the four studies done so far with CAB, including the study reported in this paper) have shown that modest query changes may substitute some papers at the citation selection

threshold, but the truly seminal papers have citations of such magnitude that they are invulnerable to modest query changes. For this reason, the cutoff threshold for citations has been, and should be, set slightly lower, to compensate for query uncertainties.

Third, there may be situations where at least minimal citation representation is desired from each of the major technical thrust areas in the documents retrieved. In this case, the retrieved documents could be clustered into the major technical thrust areas, and the CAB process could be performed additionally on the documents for each cluster. The additional references identified with the cluster-level CAB process, albeit with lower citations than from the aggregated non-clustered CAB process, would then be added to the list obtained with the aggregated CAB process. The first author has not found this cluster-level CAB process necessary for any of the four disciplines studied with CAB so far.

Fourth, there may be errors in citation counts due to references errors, and the subsequent fragmenting of a reference's occurrence frequency metric into smaller metric values. Care needs to be taken in insuring that a given reference is not fissioned into multiple large fragments, that are not subsequently combined.

How large would this fragmenting effect be? There have been a number of published studies estimating these types of data entry errors on SCI citation results (Gosling et al, 2004; Fenton et al, 2000; Putterman et al, 1991). Essentially all the articles retrieved used the same approach. They selected a sample of journal papers from a journal or journals, and compared the references against the originals. In the words of one of the retrieved papers' authors: "To evaluate the reference accuracy in the Journal of Dermatology and the Korean Journal of Dermatology, we randomly selected 100 references from each journal and checked them against the original articles." (Lee and Lee, 1999). They generated metrics for citation errors, and presented the results statistically. There was a range of results, but 'significant' errors appeared to be in the range of about ten percent.

The first author did a study in early 2003 (unpublished) examining the differences between numerical outputs in the Times Cited field in the Science Citation Index (SCI) and the Cited Reference Search capability in the SCI. This difference reflected the error in entering reference data in the

SCI, and would directly lead to fragmenting of the reference occurrence frequency metrics.

The SCI allows computation of citation counts for a paper by two different methods. One approach is the Times Cited field associated with the paper of interest (Pi). The other is the Cited Reference Search capability. The Times Cited field essentially counts links between the SCI record of the Pi and the other SCI records that contain references to Pi in their Cited References field. Any errors in how Pi is referenced in these other SCI records will nullify a link. The Cited Reference Search capability lists all references for Pi, and groups them by similarity. One group is those references that have been entered correctly, and have established the link to the Times Cited field.

Citation counts for ten highly cited papers were computed for each method. The first author's name, as it appeared in the SCI record of the actual paper, was the only variant used for the experiment. The Times Cited count averaged about four percent less than the Cited Reference Search. This appeared due to errors in entering the journal volume, page, or year. Any errors in entering the first author's name would exacerbate this under-representation. From observation, the greatest source of author name error appeared to be in the treatment of the middle initial (exclusion, if the middle initial appeared in the SCI record of the actual paper). In the study above, not all the errors made in entering data could be identified, and therefore the four percent number is a lower bound on the differential.

For statistical purposes in representing numbers of citations, the Times Cited field is adequate. For a more accurate representation, the Cited Reference Search would be required. Using a stem of the author's name (followed by wildcards) to obtain estimates of the differences due to name entry errors is very time consuming, and does not fully obviate the problem, since it is not known how the error would have impacted any stem selected. For almost any conceivable application, this additional level of complexity and time would not justify the probable slight increase in citation count accuracy.

Fifth, the CAB approach is most accurate for recent references, and its accuracy drops as the references recede into the distant past. This results from the tendency of authors to reference more recent documents and, given the restricted real estate in journals, not reference the original documents. To get better representation, and more accurate citation numbers, for early

historical documents, the more recent references need to be retrieved, collected into a database, and have their references analyzed in a similar manner (essentially examining generation of citations).

As an example of what would be required for the early historical documents, assume 150 reference documents are selected for the primary Background study, and the retrieved database is for 2001. Assume there is an average of twenty references per retrieved record for a total of 3000 references. Assume half of these references are in the SCI, for a total of 1500 references. All these 1500 references could be retrieved, could constitute the new database, the critical references in this database could be identified, and the process repeated ad infinitum. Or, to make the numbers more manageable in terms of number of iterations required, an upper limit on publication date could be specified for each succeeding iteration. Thus, for an initial retrieval of 2001 as in the example, the next retrieval could be for references prior to 1980, then the following retrieval would be for references prior to 1960. However, for most literature surveys, this iterative approach would be un-necessary, since recent references tend to be of primary interest.

Sixth, high citation frequencies are not unique to seminal documents only; different types of references can have high citation frequencies. Documents that contain critical research advances, and were readily accessible in the open literature, tend to be cited highly, and represent the foundation of the CAB approach. Application of CAB to three technical research areas so far (in addition to the present Nonlinear Dynamics study) shows that this type of document is predominant in the highly cited references list. Books or review articles also appear on the highly cited references list. These documents do not usually represent new advances, but rather are summaries of the state of the art (and its Background) at the time the document was written. These types of documents are still quite useful as Background material. Finally, documents that receive large numbers of citations highly critical of the document could be included in the list of highly cited documents. In three studies so far, the first author has not identified such papers in the detailed development of the Background.

Additionally, one of the three application studies concerns high speed compressible flow, a discipline in which the first author worked decades ago. Using the CAB approach, the first author found that all the key historical documents with which he was familiar were identified, and all the

historical documents identified appeared to be important. Thus, for that data point at least, the weaknesses identified above (imbalances, undervaluing early historical references, unwanted highly cited documents) did not materialize. To insure that any critical documents were not missed because of imbalance problems, the threshold was set a little bit lower to be more inclusive.

The converse problem to multiple types of highly cited references, some of which may not be the seminal documents desired, is influential references that do not have substantial citation frequencies. If the authors of these references did not publish them in widely and readily accessible forums, or if they do not contain appropriate verbiage for optimal query accessibility, then they might not have received large numbers of citations. Additionally, journal or book space tends to be limited, with limited space for references. In this zero-sum game for space, research authors tend to cite relatively recent records at the expense of the earlier historical records. Also, extremely recent but influential references have not had the time to accumulate sufficient citations to be listed above the selection threshold on the citation frequency table. Methods of including these influential records located at the wings of the temporal distribution will be described in the following implementation section. Inclusion of the references that were not widely available when published is more problemmatical, and tends to rely on the Background developers' personal knowledge of these documents, and their influence.

CONCEPT IMPLEMENTATION

To identify the total candidate references for the Background sub-section, a table similar in structure to Table 1, but containing all the references from the retrieved records, is constructed. A threshold frequency for selection can be determined by arbitrary inspection (i.e., a Background section consisting of 150 key references is arbitrarily selected). The first author has found a dynamic selection process more useful. In this dynamic process, references are selected, analyzed, and grouped based on their order in the citation frequency table until the resulting Background is judged sufficiently complete by the Background developers.

To insure that the influential documents at the wings of the temporal distribution are included, the following total process is used. The reference

frequency table is ordered by inverse frequency, as above, and a high value of the selection frequency threshold is selected initially. Then, the table is re-ordered chronologically. The early historical documents with citation frequencies substantially larger than those of their contemporaries are selected, as are the extremely recent documents with citation frequencies substantially larger than those of their contemporaries. By contemporaries, it is meant documents published in the same time frame, not limited to the same year. Then, the dynamic selection process defined above is applied to the early historical references, the intermediate time references (those falling under the high frequency threshold), and the extremely recent references.

Table 2 is an example of the final references that would have been selected for the Background section of the Nonlinear Dynamics study using CAB, had an extensive Background section been desired. The first reference listed, Einstein's 1917 paper, had many more citations than any papers published in the 1910s or 1920s. In fact, there were half a dozen papers published between 1831 and 1931 that had four citations each, and these were the closest to Einstein's paper. This is a graphic example of how we interpret a paper's having substantially more citations than its contemporaries.

TABLE 2 – SEMINAL DOCUMENTS SELECTED FOR INCLUSION IN BACKGROUND

AUTHOR	YEAR	SOURCE	VOL	PAGE	# CIT
EINSTEIN A	1917	VERHAND DEUT PHYS GE	V19	P82	13
LAMB H	1932	HYDRODYNAMICS			14
WIGNER E	1932	PHYS REV	V40	P749	11
KOLMOGOROV AN	1937	B MGU A	V1	P1	10
HUSIMI K	1940	P PHYS-MATH SOC JPN	V22	P264	10
GABOR D	1946	J I ELEC ENG 3	V93	P429	11
HODGKIN AL	1952	J PHYSIOL-LONDON	V117	P500	30
TURING AM	1952	PHILOS T ROY SOC B	V237	P37	27
CODDINGTON EA	1955	THEORY ORDINARY DIFF			15
ANDERSON PW	1958	PHYS REV	V109	P1492	21
FITZHUGH R	1961	BIOPHYS J	V1	P445	24
CHANDRASEKHAR S	1961	HYDRODYNAMIC HYDROMA			23
LORENZ EN	1963	J ATMOS SCI	V20	P130	115
MELNIKOV VK	1963	T MOSCOW MATH SOC	V12	P1	23
HENON M	1964	ASTRON J	V69	P73	18
SMALE S	1967	B AM MATH SOC	V73	P747	19
OSELEDEC VI	1968	T MOSCOW MATH SOC	V19	P197	25
GUTZWILLER MC	1971	J MATH PHYS	V12	P343	42
RUELLE D	1971	COMMUN MATH PHYS	V20	P167	23
ZAKHAROV VE	1972	SOV PHYS JETP-USSR	V34	P62	21

NAYFEH AH	1973	PERTURBATION METHODS			24
HENON M	1976	COMMUN MATH PHYS	V50	P69	41
ROSSLER OE	1976	PHYS LETT A	V57	P397	39
MAY RM	1976	NATURE	V261	P459	35
BENETTIN G	1976	PHYS REV A	V14	P2338	27
MACKEY MC	1977	SCIENCE	V197	P287	35
NICOLIS G	1977	SELF ORG NONEQUILIBR			26
FEIGENBAUM MJ	1978	J STAT PHYS	V19	P25	28
NAYFEH AH	1979	NONLINEAR OSCILLATIO			62
CHIRIKOV BV	1979	PHYS REP	V52	P263	43
PACKARD NH	1980	PHYS REV LETT	V45	P712	54
LANG R	1980	IEEE J QUANTUM ELECT	V16	P347	29
WINFREE AT	1980	GEOMETRY BIOL TIME			25
TAKENS F	1981	LECT NOTES MATH	V898	P366	97
BRODY TA	1981	REV MOD PHYS	V53	P385	35
HOPFIELD JJ	1982	P NATL ACAD SCI-BIOL	V79	P2554	37
GUCKENHEIMER J	1983	NONLINEAR OSCILLATIO			149
GRASSBERGER P	1983	PHYSICA D	V9	P189	94
GRASSBERGER P	1983	PHYS REV LETT	V50	P346	76
FUJISAKA H	1983	PROG THEOR PHYS	V69	P32	61
GREBOGI C	1983	PHYSICA D	V7	P181	26
BOHIGAS O	1984	PHYS REV LETT	V52	P1	54
KURAMOTO Y	1984	CHEM OSCILLATIONS WA			49
HELLER EJ	1984	PHYS REV LETT	V53	P1515	44
AREF H	1984	J FLUID MECH	V143	P1	29
WOLF A	1985	PHYSICA D	V16	P285	103
ECKMANN JP	1985	REV MOD PHYS	V57	P617	75
BERRY MV	1985	P ROY SOC LOND A MAT	V400	P229	35
MILNOR J	1985	COMMUN MATH PHYS	V99	P177	28
FRASER AM	1986	PHYS REV A	V33	P1134	49
THEILER J	1986	PHYS REV A	V34	P2427	34
BROOMHEAD DS	1986	PHYSICA D	V20	P217	26
FARMER JD	1987	PHYS REV LETT	V59	P845	36
SKARDA CA	1987	BEHAV BRAIN SCI	V10	P161	25
TEMAM R	1988	INFINITE DIMENSIONAL			31
PARKER TS	1989	PRACTICAL NUMERICAL			40
OTTINO JM	1989	KINEMATICS MIXING ST			35
CASDAGLI M	1989	PHYSICA D	V35	P335	32
OSBORNE AR	1989	PHYSICA D	V35	P357	25
PECORA LM	1990	PHYS REV LETT	V64	P821	177
OTT E	1990	PHYS REV LETT	V64	P1196	142
GUTZWILLER MC	1990	CHAOS CLASSICAL QUAN			88
WIGGINS S	1990	INTRO APPL NONLINEAR			61
SUGIHARA G	1990	NATURE	V344	P734	35
KANEKO K	1990	PHYSICA D	V41	P137	30
AIHARA K	1990	PHYS LETT A	V144	P333	30
DITTO WL	1990	PHYS REV LETT	V65	P3211	29
MEHTA ML	1991	RANDOM MATRICES			51
SAUER T	1991	J STAT PHYS	V65	P579	48
PECORA LM	1991	PHYS REV A	V44	P2374	29
HUNT ER	1991	PHYS REV LETT	V67	P1953	28
THEILER J	1992	PHYSICA D	V58	P77	66

PYRAGAS K	1992	PHYS LETT A	V170	P421	59
LICHTENBERG AJ	1992	REGULAR CHAOTIC DYNA			58
KENNEL MB	1992	PHYS REV A	V45	P3403	33
KOCAREV L	1992	INT J BIFURCAT CHAOS	V2	P709	31
PRESS WH	1992	NUMERICAL RECIPES C			29
GARFINKEL A	1992	SCIENCE	V257	P1230	27
MARCUS CM	1992	PHYS REV LETT	V69	P506	26
ALEXANDER JC	1992	INT J BIFURCAT CHAOS	V2	P795	25
CROSS MC	1993	REV MOD PHYS	V65	P851	105
OTT E	1993	CHAOS DYNAMICAL SYST			97
CUOMO KM	1993	PHYS REV LETT	V71	P65	57
ABARBANEL HDI	1993	REV MOD PHYS	V65	P1331	54
PLATT N	1993	PHYS REV LETT	V70	P279	38
CUOMO KM	1993	IEEE T CIRCUITS-II	V40	P626	34
WU CW	1993	INT J BIFURCAT CHAOS	V3	P1619	28
HEAGY JF	1994	PHYS REV E	V50	P1874	40
OTT E	1994	PHYS LETT A	V188	P39	40
STROGATZ SH	1994	NONLINEAR DYNAMICS C			35
ASHWIN P	1994	PHYS LETT A	V193	P126	33
LASOTA A	1994	CHAOS FRACTALS NOISE			30
HEAGY JF	1994	PHYS REV E	V49	P1140	30
ROY R	1994	PHYS REV LETT	V72	P2009	28
SCHIFF SJ	1994	NATURE	V370	P615	28
RULKOV NF	1995	PHYS REV E	V51	P980	59
NAYFEH AH	1995	APPL NONLINEAR DYNAM			46
KOCAREV L	1995	PHYS REV LETT	V74	P5028	40
KATOK A	1995	INTRO MODERN THEORY			27
ROSENBLUM MG	1996	PHYS REV LETT	V76	P1804	77
ABARBANEL HDI	1996	ANAL OBSERVED CHAOTI			45
KOCAREV L	1996	PHYS REV LETT	V76	P1816	38
LAI YC	1996	PHYS REV LETT	V77	P55	27
ASHWIN P	1996	NONLINEARITY	V9	P703	27
ZELEVINSKY V	1996	PHYS REP	V276	P85	26
KANTZ H	1997	NONLINEAR TIME SERIE			54
PIKOVSKY AS	1997	PHYSICA D	V104	P219	43
PECORA LM	1997	CHAOS	V7	P520	40
ROSENBLUM MG	1997	PHYS REV LETT	V78	P4193	39
BEENAKKER CWJ	1997	REV MOD PHYS	V69	P731	25
GAMMAITONI L	1998	REV MOD PHYS	V70	P223	52
GUHR T	1998	PHYS REP	V299	P189	37
VANWIGGEREN GD	1998	SCIENCE	V279	P1198	32
GOEDGEBUER JP	1998	PHYS REV LETT	V80	P2249	29
TASS P	1998	PHYS REV LETT	V81	P3291	29
HEGGER R	1999	CHAOS	V9	P413	27
FISCHER I	2000	PHYS REV A	V6201	P1801	16
MATEOS JL	2000	PHYS REV LETT	V84	P258	15
WANG W	2000	CHAOS	V10	P248	14
VANAG VK	2000	NATURE	V406	P389	13

These results were examined by the authors. They judged that all papers in the table were relevant for a Background section, or review paper. Some of

the earliest papers (e.g., Wigner or Anderson) are concerned with random systems and not with chaotic systems, but the methods they employed influenced how to view and contrast with chaotic systems mathematically.

They also identified about 6% additional papers that he would have included in a Background section. These papers tended to have relatively high total citations, but relatively low citations from the Nonlinear Dynamics papers in the present database. Some of the papers omitted were straight plasma physics focused on nuclear fusion tokomak physics. The system was naturally very Nonlinear so the work involved Nonlinear Dynamics, but the purpose of the paper was fusion and not advancing the field of Nonlinear Dynamics. This could cause Nonlinear Dynamics authors not to reference these papers widely. Their references come from the plasma community. Finally, some papers are highly cited, but then get replaced by better (or more easily read) papers by the same author. The newer citations tend to cite the author's newer paper.

The analysis and discussion above have focused on the contents of the Background; i.e., which documents should be included. In some cases, the Abstracts of the seminal references have been retrieved and clustered, to produce a structure for the Background. Thus, the CAB approach can be used to determine both the content and structure of the Background section. Again, CAB does not exclude content and structure determinations by the experts. CAB can be viewed as the starting point for content and structure determination, upon which the experts can build with their own insights and experience.

While the CAB approach is systematic, it is not automatic. Judgment is required to determine when an adequate number of references has been selected for the Background, and further judgement is required to analyze, group, and link the references to form a cohesive Background section. Additionally, the highly influential references that were not highly cited due to insufficient dissemination should be included by the Background developers, if they know of such documents. CAB is not meant to replace individual judgment or specification of Background material. CAB is meant to augment individual judgment and reference selection, as reflected in its name of Citation-Assisted.

CONCLUSIONS

A method for systematically determining seminal references for inclusion in literature surveys or Background sections of research documents has been described. It is based on the assumption that seminal documents tend to be highly cited. CAB is being applied presently to three applications studies, and the results so far are much superior to those used by the first author for background development in any other study.

REFERENCES FOR SECTION 11

Calne DB, Calne R. Citation of original research. Lancet**.** 340 (8813): 244-244. Jul 25 1992.

Della Mea V, Mizzaro S. Measuring retrieval effectiveness: A new proposal and a first experimental validation. Journal of the American Society for Information Science and Technology. 55 (6): 530-543. Apr 2004.

Fenton JE, Brazier H, De Souza A, Hughes JP, Mcshane DP. The accuracy of citation and quotation in otolaryngology/head and neck surgery journals. Clinical Otolaryngology. 25 (1): 40-44. Feb 2000.

Gosling CM, Cameron M, Gibbons PF. Referencing and quotation accuracy in four manual therapy journals. Manual Therapy. 9 (1): 36-40 Feb 2004.

Harter SP, Hert CA. Evaluation of information retrieval systems: Approaches, issues, and methods. Annual Review of Information Science and Technology. 32: 3-94 1997.

Kagolovsky Y, Moehr JR. A new look at information retrieval evaluation: Proposal for solutions. Journal of Medical Systems. 28 (1): 103-116. Feb 2004.

Kagolovsky Y, Moehr JR. Evaluation of information retrieval. Journal of Medical Systems. In Press.

Kantor PB. Information-retrieval techniques. Annual Review of Information Science and Technology. 29: 53-90. 1994.

Khan MS, Khor S. Enhanced Web document retrieval using automatic query expansion**.** Journal of the American Society for Information Science And Technology. 55 (1): 29-40. Jan 1 2004.

Kostoff, R. N., Shlesinger, M., and Tshiteya, R. "Nonlinear Dynamics Roadmaps using Bibliometrics and Database Tomography". International Journal of Bifurcation and Chaos. January 2004.

Lee SY, Lee JS. A survey of reference accuracy in two Asian dermatologic journals (the Journal of Dermatology and the Korean Journal of

Dermatology). International Journal of Dermatology. 38 (5): 357-360. May 1999.
Liu, M.X . Progress In Documentation - The Complexities of Citation Practice – A Review of Citation Studies. Journal of Documentation. 49 (4): 370-408 Dec 1993
Macroberts, M.H, Macroberts B.R. Problems of Citation Analysis - A Critical-Review. Journal of the American Society for Information Science. 40 (5): 342-349. 1989.
MacRoberts, M.H, MacRoberts, B.R Citation content analysis of a botany journal. Journal of the American Society for Information Science. 48 (3): 274-275 Mar 1997
MacRoberts, M.H,, MacRoberts, B.R. Problems of citation analysis. Scientometrics. 36 (3): 435-444 Jul-Aug 1996.
Moravcsik MJ, Murugesan P. Some results on function and quality of citations. Social Studies of Science. 5 (1): 86-92. 1975.
Putterman C, Lossos IS. Author, verify your references - or, the accuracy of references in Israeli medical journals. Israel Journal of Medical Sciences. 27 (2): 109-112. Feb 1991.
Shadish WR, Tolliver D, Gray M, Sengupta SK Author judgments about works they cite - 3 studies from psychology journals. Social Studies of Science. 25 (3): 477-498. Aug 1995.
Smith, A.J, Goodman, N.W. The hypertensive response to intubation. Do researchers acknowledge previous work? Canadian Journal of Anaesthesia-Journal Canadien D Anesthesie. 44 (1): 9-13 Jan 1997

Section 12. The Difference between Highly and Poorly Cited Medical Articles in the Journal Lancet

(based on Kostoff, R.N. "The Difference between Highly and Poorly Cited Medical Articles in the Journal Lancet". Scientometrics. 72: 3. 513–520. 2007)

OVERVIEW

Characteristics of highly and poorly cited research articles (with Abstracts) published in The Lancet over a three-year period were examined. These characteristics included numerical (numbers of authors, references, citations, Abstract words, journal pages), organizational (first author country, institution type, institution name), and medical (medical condition, study approach, study type, sample size, study outcome). Compared to the least cited articles, the most cited have three to five times the median number of authors per article, fifty to six hundred percent greater median number of references per article, 110 to 490 times the median number of citations per article, 2.5 to almost seven times the median number of Abstract words per article, and 2.5 to 3.5 times the median number of pages per article.

The most cited articles' medical themes emphasize breast cancer, diabetes, coronary circulation, and HIV immune system problems, focusing on large-scale clinical trials of drugs. The least cited articles' themes essentially don't address the above medical issues, especially from a clinical trials perspective, cover a much broader range of topics, and have much more emphasis on social and reproductive health issues. Finally, for sample sizes of clinical trials specifically, those of the most cited articles ranged from a median of about 1500 to 2500, whereas those of the least cited articles ranged from 30 to 40.

KEYWORDS

Citation Analysis, Research Impact, The Lancet, Clinical Trials, Drugs, Text Mining; Bibliometrics,; Epidemiology; Meta-Analyses.

BACKGROUND

Medical researchers who publish want their papers to be highly cited. Three necessary conditions for highly cited papers under control of the research

author(s) are high intrinsic quality, high research activity discipline, and high circulation journal (1-5) The present section examines other characteristics of highly cited medical papers, to ascertain unique attribute patterns in these papers. It extends a comparative method used by the author in recent studies, whereby the attributes of the most cited papers are compared with those of the least cited papers (e.g., 6). This contrast between most and least has been shown to delineate differences between the two groups dramatically, for single discipline studies. It was desired to extend the technique to a more heterogeneous discipline (general medical studies), and use The Lancet, a British-based medical journal, as the test-bed. The Lancet is a well-regarded leading medical journal, and contains a substantial number of very highly cited papers.

APPROACH

A database of Lancet papers published in a narrow time window (for time normalization) and accessed through the Science Citation Index (SCI) was generated, and the detailed attributes (characteristics) of most and least cited papers were identified. Specifically, all documents classified by the SCI as articles and published in Lancet from 1997-1999 were examined initially. The majority of records classified as Articles (~2/3) did not have Abstracts. The articles with Abstracts are viewed as complete research articles, and are analyzed in more detail. Characteristics evaluated for each of the most and least cited articles are based on the author's recent analyses of most and least cited articles in neuro-psychology (6) and desalination journals, and the anthrax discipline, with the exception that a wider range of characteristics was used for the present analysis.

Two key issues in the approach are the selection of citations as a key metric, and the quantification of citations used to define 'high' and 'low'. Citations are used as a metric for research impact/ quality, and allocation of papers into high and low citation categories is a proxy for separation into high and low quality categories. However, for clinical trials specifically, some caveats are required. As Ioannidis states (7), in an analysis of highly cited clinical trials, "Of 49 highly cited original clinical research studies, 45 claimed that the intervention was effective. Of these, 7 (16%) were **contradicted** by subsequent studies, 7 others (16%) had found effects that were stronger than those of subsequent studies, 20 (44%) were replicated, and 11 (24%) remained largely unchallenged."

The second key issue is the quantitative determination of most and least cited. If too few articles are selected (e.g., the top or bottom two or three, or 1%), the statistics will be insufficient to allow for general conclusions. If too many articles are selected (e.g., the top or bottom 40 or 50, or 20%), the gradations within each category will be too large for 'most' and 'least' to have meaning. Based on previous studies, and on tests where the number/ fraction of articles in 'most' and 'least' categories were examined, the top and bottom 5% of articles with Abstracts were selected for each year, and their characteristics compared. The number of articles in each category was kept constant at seventeen per year.

RESULTS

Tables 1 and 2 list detailed characteristics of the most cited and the least cited research articles for 1998. Tables 1A and 1B contain the bibliometrics, and Tables 2A and 2B contain the medical/ technical issues.

In Tables 1A and 1B, starting from the second from left column, the characteristics listed are number of authors, number of references, number of citations as shown in the Times Cited field in the Science Citation Index, number of Abstract words, number of pages, reprint author country, reprint author institution type, and reprint author institution name. The reprint author is typically the first author. In Tables 2A and 2B, starting from the second from left column, the characteristics listed are medical condition/ theme, approach, study type, and sample size.

TABLE 1A – MOST CITED - BIBLIOMETRICS

1998 MOST CITED ARTICLES

ART #	#AUTH	#REFS	#CITES	#ABS WORDS	#PAGES	1ST AUTH COUNTRY	1ST AUTH INST TYPE	1ST AUTH INST NAME
1	400	47	2799	529	17	ENGLAND	HOSP	RADCLIFFE
2	10	33	1563	438	8	SWEDEN	UNIV	UPPSALA
3	6	27	1232	606	17	ENGLAND	HOSP	RADCLIFFE
4	11	27	1014	389	7	FRANCE	HOSP	PITIE SALPETRIERE
5	13	20	895	423	12	ENGLAND	HOSP	RADCLIFFE
6	188	30	782	239	6	USA	HOSP	CLEVELAND CLIN
7	25	24	730	286	8	NEW ZEALAND	UNIV	WELLINGTON
8	326	19	650	579	13	ENGLAND	HOSP	RADCLIFFE
9	9	36	548	277	4	USA	HOSP	BRIGHAM/ WOMENS

10	5	32	507	323	7	ITALY	HOSP	SALVATORE MAUGERI
11	5	8	484	328	9	SCOTLAND	HOSP	WESTERN GEN
12	4	25	480	166	3	AUSTRALIA	HOSP	ST. VINCENTS
13	13	19	463	262	7	GERMANY	UNIV	HEIDELBERG
14	119	39	447	296	7	CANADA	HOSP	LONDON HEALTH
15	9	18	412	307	4	ENGLAND	HOSP	ROYAL MARSDEN
16	8	20	408	341	5	ITALY	HOSP	INST ONCOLOGY
17	5	33	407	288	5	USA	HOSP	BRIGHAM/WOMENS

TABLE 1B – LEAST CITED - BIBLIOMETRICS

1998 LEAST CITED ARTICLES

ART #	#AUTH	#REFS	#CITES	#ABS WORDS	#PAGES	1ST AUTH COUNTRY	1ST AUTH INST TYPE	1ST AUTH INST NAME
314	1	19	0	68	3	SCOTLAND	UNIV	GLASGOW
313	2	30	0	202	5	NETHERLANDS	UNIV	UTRECHT
312	2	37	0	80	5	CANADA	UNIV	TORONTO
311	2	35	2	111	3	USA	HOSP	TEXAS ARRYTHMIA
310	1	13	4	124	4	ENGLAND	UNIV	LONDON
309	6	15	4	266	3	AUSTRIA	HOSP	WILHELMINENSPITAL
308	3	4	5	198	1	ENGLAND	HOSP	CHARING CROSS
307	1	15	5	143	5	ENGLAND	FOUND	GLOBAL FORUM
306	2	16	5	127	4	NORWAY	INST	DIAKONHJEMMETS
305	2	6	6	154	3	USA	UNIV	CORNELL
304	8	13	6	213	3	AUSTRALIA	GOVT	ACHS
303	3	27	6	297	5	SCOTLAND	HOSP	WESTERN GEN
302	3	43	8	62	5	WALES	UNIV	CARDIFF
301	1	55	8	72	6	USA	UNIV	ST. LOUIS
300	10	0	9	88	2	ENGLAND	FOUND	AIDS CONSORTIUM
299	6	12	9	175	3	ENGLAND	UNIV	NOTTINGHAM
298	3	13	9	314	3	FRANCE	UNIV	REIMS

TABLE 2A – MOST CITED – MEDICAL ISSUES

1998 MOST CITED ARTICLES

ART #	MEDICAL CONDITION/THEME	APPROACH	STUDY TYPE	SAMPLE SIZE	OUTCOME
1	DIABETES	DRUGS	CLINICAL TRIAL	LARGE(3867)	PARTIALLY FAVORABLE
2	HYPERTENSION	DRUGS	CLINICAL TRIAL	LARGE(18790)	FAVORABLE
3	BREAST CANCER	DRUGS	META-ANALYSIS	LARGE(37000)	PARTIALLY FAVORABLE
4	HEPATITIS VIRUS	DRUGS	CLINICAL TRIAL	LARGE(832)	FAVORABLE
5	DIABETES	DRUGS	CLINICAL TRIAL	LARGE(2241)	MODERATELY FAVORABLE
6	CORONARY STENT	DRUGS	CLINICAL	LARGE(2399)	FAVORABLE

			TRIAL		
7	ALLERGIES	QUESTIONNAIRES	EPIDEMIOLOGY	LARGE (463801)	FAVORABLE
8	BREAST CANCER	DRUGS	CLINICAL TRIAL	LARGE (30000)	MODERATELY FAVORABLE
9	BREAST CANCER	BLOOD ASSAYS	CLINICAL ASSAY	LARGE (1017)	FAVORABLE
10	MYOCARDIAL INFARCT	HEART MONITOR	CLINICAL TRIAL	LARGE (1284)	FAVORABLE
11	CAROTID STENOSIS	SURGICAL	CLINICAL TRIAL	LARGE (3024)	MODERATELY FAVORABLE
12	HIV-1	MECHANISM MODEL	THEORETICAL	N.A.	HYPOTHESIS
13	STROKE	DRUGS	CLINICAL TRIAL	LARGE (800)	MARGINALLY FAVORABLE
14	MULTIPLE SCLEROSIS	DRUGS	CLINICAL TRIAL	LARGE (533)	FAVORABLE
15	BREAST CANCER	DRUGS	CLINICAL TRIAL	LARGE (2494)	INCONCLUSIVE
16	BREAST CANCER	DRUGS	CLINICAL TRIAL	LARGE (5408)	INCONCLUSIVE
17	MYOCARDIAL INFARCT	BLOOD ASSAYS	CLINICAL ASSAY	LARGE (14916)	FAVORABLE

TABLE 2B – LEAST CITED – MEDICAL ISSUES

1998 LEAST CITED ARTICLES

ART #	MEDICAL CONDITION/THEME	APPROACH	STUDY TYPE	SAMPLE SIZE	OUTCOME
314	STROKE	TREATMENT PROTOCOL	ASSESSMENT	N.A.	N.A.
313	STROKE	TREATMENT PROTOCOL	ASSESSMENT	N.A.	N.A.
312	STROKE	TREATMENT PROTOCOL	ASSESSMENT	N.A.	N.A.
311	ARRYTHMIA	TREATMENT HYPOTHESIS	THEORETICAL	N.A.	N.A.
310	GLOBAL HEALTH	EVALUATION	PROPOSE REFORMS	N.A.	N.A.
309	TORTURE INJURIES	BONE SCANS	CLINICAL TRIAL	SMALL(50)	FAVORABLE
308	PULMONARY EMBOLISM	DRUGS	CASE STUDY	SMALL(1)	UNFAVORABLE
307	WHO' IMPACT	EVALUATION	PROPOSE REFORMS	N.A.	N.A.
306	SOCIAL IMPACT	EVALUATION	PROPOSE REFORMS	N.A.	N.A.
305	FOOD AID	EVALUATION	ASSESSMENT	N.A.	N.A.
304	PSITTACOSIS	EPIDEMIOLOGY	CLINICAL DIAGNOSTICS	SMALL(16)	FAVORABLE
303	BRAIN TUMORS	SURGICAL	CLINICAL TRIAL	SMALL(40)	FAVORABLE
302	DATA INTERPRETATION	TEACHING	PROTOCOLS	N.A.	N.A.
301	ACID-BASE DISORDERS	DIAGNOSTIC TESTS	CLINICAL DIAGNOSTICS	N.A.	N.A.
300	HIV	TREATMENT ACCESS	PROPOSE REFORMS	N.A.	N.A.
299	PLACENTAL PERFUSION	DIAGNOSTIC TESTS	CLINICAL TRIAL	SMALL(15)	FAVORABLE
298	FERTILIZATION	QUESTIONNAIRE	SURVEY	SMALL(48)	INCONCLUSIVE

Table 3 summarizes the bibliometrics results for 1997-1999 for most and least cited articles, while Table 4 summarizes the medical issues results for the same records. The numbers in parentheses after the text entries reflect the number of occurrences. Only those entries with occurrences greater than one are shown.

TABLE 3 – 1997-1999 – MOST/ LEAST CITED – BIBLIOMETRICS - SUMMARY

ART #	#AUTH	#REFS	#CITES	#ABS WORDS	#PAGES	1ST AUTH COUNTRY	1ST AUTH INST TYPE	1ST AUTH INST NAME
1997 SUMMARY - MOST CITED								
AVERAGE	46.59	30.82	692.1	358.8	7.059	W. EUR (10)	UNIV (9)	HARVARD (3)
MEDIAN	9	29	660	336	7	N. AMERICA (5)	HOSP (5)	
STD DEV	75.96	10.68	220.2	96.33	2.703		INST (3)	
1997 SUMMARY - LEAST CITED								
AVERAGE	3.118	21.06	5.118	178.1	3.529	W. EUROPE (9)	UNIV (7)	
MEDIAN	2	20	5	142	3	N. AMERICA (5)	HOSP (5)	
STD DEV	2.369	10.56	3.257	105	1.281		INST (3)	
1998 SUMMARY - MOST CITED								
AVERAGE	68	26.88	813	357.5	8.176	WEST EUR(11)	UNIV(3)	RADCLIFFE(4)
MEDIAN	10	27	548	323	7	N. AMERICA(4)	HOSP(14)	BRIGHAM/WOMENS(2)
STD DEV	122.1	9.393	606.9	121.7	4.231	AUSTRALASIA(2)		
1998 SUMMARY - LEAST CITED								
AVERAGE	20.0	10.9	209.0	118.7	2.6	WEST EUR(12)	UNIV(9)	
MEDIAN	2	15	5	143	3	N. AMER(4)	HOSP(4)	
STD DEV	2.6	14.9	3.1	80.0	1.3		FOUND(2)	
1999 SUMMARY - MOST CITED								
AVERAGE	32.83	19.21	342.5	207.3	4.506	WEST EUR (14)	UNIV (9)	UPPSALA (3)
MEDIAN	11	29	489	258	7	N. AMERICA (2)	HOSP (6)	
STD. DEV.	40.29	8.276	323.4	110.8	2.352		INST (2)	
1999 SUMMARY - LEAST CITED								
AVERAGE	30.51	18.88	316.4	197.5	4.26	W. EUR. (10)	UNIV (12)	
MEDIAN	4	5	1	38	2	AUSTRALASIA (3)	HOSP (2)	
STD. DEV.	37.72	7.65	309.9	107.1	2.323	N. AMERICA (2)	INST (3)	

TABLE 4 – 1997-1999 – MOST/ LEAST CITED – MEDICAL ISSUES - SUMMARY

ART #	MEDICAL CONDITION/THEME	APPROACH	STUDY TYPE	SAMPLE SIZE	OUTCOME
1997 SUMMARY - MOST CITED					
	CORONARY (6)	DRUGS (8)	CLIN TRIAL (8)	LARGE	FAVORABLE (12)
	MORTALITY (3)	DATA ANALYSIS (4)	LAB TESTS (5)	MEDIAN (1734)	MODERATELY FAV (4)
	BREAST CANCER (2)	BIOPSY (2)	EPIDEMIOL (4)		
		LAB TESTS (2)			
1997 SUMMARY - LEAST CITED					
	RADIOLOGY (4)	OVERVIEW (5)	OVERVIEW (6)	LARGE (4)	FAVORABLE (5)
		LAB TESTS (3)	ASSESSMENTS (2)	MEDIAN (755)	

	ANALYSIS (2)	EPIDEMIOLOGY (2)	SMALL (2)	
	SURVEY (2)		MEDIAN (7)	
1998 SUMMARY - MOST CITED				
BREAST CANCER(5)	DRUGS(11)	CLINICAL TRIAL(12)	LARGE(16)	FAVORABLE(8)
DIABETES(2)	BLOOD ASSAYS(2)	CLINICAL ASSAYS(2)	(MEDIAN 3445)	MODERATELY FAVOR(3)
MYOCARDIAL INFARCT(2)				PARTIALLY FAVOR(2)
1998 SUMMARY - LEAST CITED				
SOCIAL HEALTH(4)	EVALUATION(4)	ASSESSMENT(4)	SMALL(6)	N.A.(11)
STROKE(3)	TREATMENT PROTOCOL(3)	PROPOSE REFORM(4)	(MEDIAN 28)	FAVORABLE(4)
	DIAGNOSTIC TESTS(2)	CLINICAL TRIAL(3)		UNFAVORABLE(1)
		CLINICAL DIAGNOSTICS(2)		INCONCLUSIVE(1)
1999 SUMMARY - MOST CITED				
CARDIOVASCULAR(6)	DRUGS(9)	CLIN TRIAL(14)	LARGE(15)	FAVORABLE(9)
HIV-1(5)	ASSESSMENT(3)		MEDIAN(1000)	INCONCLUSIVE(4)
HYPERTENSION(2)				MODERATELY FAV(3)
1999 SUMMARY - LEAST CITED				
	LAB TEST (6)	CASE STUDY (4)	SMALL (7)	FAVORABLE (9)
	DATA ANALYSIS (4)	LAB TEST (4)	LARGE (3)	MODERATELY FAVORABLE (2)
	EVALUATION (2)	ASSESSMENT (4)		
	SURVEY (2)	SURVEY (2)		

For all three years, there are substantial differences between the most and least cited articles in both the bibliometrics and medical/ technical issues. Compared to the least cited articles, the most cited article bibliometrics reflect, for the three years, approximately:

- Fifteen to twenty times the average number of authors per article, and three to five times the median number of authors per article (see Reference 8 for a network-based depiction of numbers of authors vs numbers of citations);
- Thirty three to three hundred percent greater average number of references per article, and fifty to six hundred percent greater median number of references per article (see Reference 9 for relationshio of number of references to citations);
- One hundred and forty to one thousand times the average number of citations per article, and 110 to 490 times the median number of citations per article;
- Two to four times the average number of Abstract words per article, and 2.5 to almost seven times the median number of Abstract words per article;
- Two and a half to 3.5 times the average number of pages per article, and 2.5 to 3.5 times the median number of pages per article.

- There are no significant regional distribution differences between the most and least cited articles (only those regions with more than one article are listed);
- While the least cited articles tend to be predominately from universities, the most cited are mixed. The significance of this result is muted, because many of the hospitals tended to be part of a university, thereby blurring the university/ hospital distinction;
- Over the three year period, four institutions were represented three or more times in the most cited category (Radcliffe – 5; Uppsala – 4; Harvard – 3; Salpetriere – 3), while no institutions were represented three or more times in the least cited category.

The most cited articles' medical themes emphasize breast cancer, diabetes, coronary circulation, and HIV immune system problems, focusing on large scale clinical trials of drugs. The 1997 most cited articles were skewed by three publications from a large scale epidemiological study (Burden of Disease), thereby reducing the dominant fraction of clinical drug trials that characterized the other two years. The least cited articles' themes essentially don't address the above medical issues, especially from a clinical trials perspective, cover a much broader range of topics, and have much more emphasis on social and reproductive health issues. Because of the different nature of the least cited articles' studies relative to the most cited, many of the least cited articles' outcomes were non-applicable in the same sense as that of a drug clinical trial outcome. The favorability or unfavorability of results (assessed, but not included in tables) did not seem to be a major factor differentiating the most and least cited articles.

Median sample sizes of the most cited articles tend to be large, ranging from 1000 to 3500, while sample sizes of the least cited articles tend to be a couple of orders of magnitude smaller. However, 1997 did contain four of the least cited articles with a median sample size of 755, but none of these samples were from clinical drug trials. For sample sizes of clinical trials specifically, those of the most cited articles ranged from a median of about 1500 to 2500, whereas those of the least cited articles ranged from 30 to 40. References 10-11 provide further evidence that higher citation rates are correlated with larger clinical trial sample sizes, but exceptions can be found.

DISCUSSION AND CONCLUSIONS

Medicine has many facets, including adequacy and affordability of health care, disease and injury prevention, public health education, lab research and clinical trials, theory and experiment, individual and global health issues, and epidemiology. Out of all these possibilities that could be of substantial interest to the medical research community, the Lancet readership community has chosen to emphasize high citations to large-scale clinical drug trials on breast cancer, diabetes, coronary circulation, and immune system problems, reported by many authors in long well-referenced papers, for the time period chosen.

REFERENCES FOR SECTION 12.

1. Lehrl S. The citation frequency for prominent researchers in German otorhinolaryngology over 10 years. HNO. 53 (5): 415-422 2005.
2. Smith SD. Is an article in a top journal a top article? Financial Management. 33 (4): 133-149 2004.
3. Lee KP, Schotland M, Bacchetti P, Bero LA. Association of journal quality indicators with methodological quality of clinical research articles. JAMA-Journal of the American Medical Association. 287 (21): 2805-2808 2002.
4. Kostoff, RN. The use and misuse of citation analysis in research evaluation. Scientometrics. 43:1. September 1998.
5. Callaham M, Wears RL, Weber E. Journal prestige, publication bias, and other characteristics associated with citation of published studies in peer-reviewed journals. JAMA-Journal of the American Medical Association 287 (21): 2847-2850 2002
6. Kostoff RN, Buchtel H., Andrews J, Pfeil K. The hidden structure of neuropsychology: text mining of the journal *Cortex*: 1991-2001. Cortex. 41:2. 103-115. April 2005.
7. Ioannidis JPA. Contradicted and initially stronger effects in highly cited clinical research. JAMA-Journal of the American Medical Association 294 (2): 218-228. 2005.
8. Borner K, Dall'Asta L, Ke WM, et al. Studying the emerging global brain: Analyzing and visualizing the impact of co-authorship teams. Complexity 10 (4): 57-67 2005.
9. Schloegl C, Stock WG Impact and relevance of LIS journals: A scientometric analysis of international and German-language LIS journals -

Citation analysis versus reader survey. Journal of the American Society for Information Science and Technology 55 (13): 1155-1168 2004.
10. Gluud LL, Sorensen TIA, Gotzsche PC, et al. The journal impact factor as a predictor of trial quality and outcomes: Cohort study of hepatobiliary randomized clinical trials. American Journal of Gastroenterology 100 (11): 2431-2435 2005.
11. Easterbrook PJ, Berlin JA, Gopalan R, et al. Publication Bias In Clinical Research. Lancet 337 (8746): 867-872 1991.

Section 13. Citation Comparisons of the Journal Cortex.

(based on Kostoff, R. N., Buchtel, H., Andrews, J., and Pfeil, K. "The hidden structure of neuropsychology: Text Mining of the Journal *Cortex*: 1991-2001". Cortex. 41:2. 103-115. April 2005.)

BACKGROUND

The purpose of this section is to compare citations among papers published in three related neuropsychology journals, *Cortex*, *Neuropsychologia*, and *Brain*.

APPROACH

The following experiment was run. All articles published in *Cortex*, *Neuropsychologia*, and *Brain* in the years 1998-1999 were retrieved from SCI. There were 110 *Cortex* articles, 278 *Neuropsychologia* articles, and 341 *Brain* articles. Then, the ten most cited articles from each retrieval (the citations from each paper used for the tabulation of most and least cited are those listed in the SCI Times Cited field, and are the total citations received by each paper from all other papers in the SCI) were extracted, as well as the ten least cited articles, and various characteristics compared. The results are shown in Table 1

RESULTS

Table 1 – Comparison of most cited/ least cited papers, published 1998-1999.

	CORTEX		***NEUROPSYCHOLOGIA***		***BRAIN***	
	MOST CITED	LEAST CITED	MOST CITED	LEAST CITED	MOST CITED	LEAST CITED
# AUTH						
Average	3.9	2.8	5.2	2.6	7.1	4.6
Median	4	3	5	1	7.5	4.5
# REFS						
Average	46.3	28	52.5	26.8	68.3	42.4
Median	49	29.5	49	26	62.5	35
# CITES						
Average	21	0.8	71.3	0	166.8	2.8
median	18.5	1	67.5	0	157	3
ORG						
Institution	5	4	2	4	8	2
University	5	6	8	6	2	8
COUNTRY	4 Italy	2 Italy	4 UK	5 USA	5 UK	3 Japan

	3 France 1 Austria 1 Belgium 1 Germany	2 USA 2 Germany 2 Japan 1 Neth 1 Australia	4 USA 1 Italy 1 Canada	2 Italy 1 NZ 1 Neth 1 Australia	2 USA 2 Canada 1 Germany	1 USA 1 UK 1 France 1 Italy 1 Canada 1 Germany 1 Neth
TYPE						
Behavior	8		4			
Surgery			1		2	
Diagnostic-NI	2		5		7	
Diagnostic-INV					1	

CODE: TYPE
BEHAV=CLINICAL BEHAVIOR STUDIES
SURGERY=SURGICAL INTERVENTIONS
DIAG-NI=NON-INVASIVE DIAGNOSTIC TESTS
DIAG-INV=INVASIVE DIAGNOSTIC TESTS

CONCLUSIONS

A number of interesting observations may be made from Table 7. First, the most cited articles in *Neuropsychologia* are cited, on average, more than three times as often as the most cited articles in *Cortex*, and the most cited articles in *Brain* are cited, on average, more than twice as often as the most cited articles in *Neuropsychologia*.

Second, the most cited papers have more authors than the least cited, in all three journals, and the effect is most pronounced in *Neuropsychologia*. Additionally, the average number of authors increases with the average number of citations, ranging from about four authors of the most cited *Cortex* papers to about seven authors of the most cited *Brain* papers.

Third, the most cited papers have substantially more references than the least cited, in both journals, and the effect is most pronounced in *Neuropsychologia*. Additionally, the average number of citations increases with the average number of references (an effect observed by the first author in recent unpublished text mining studies), ranging from about 46 references in the most cited *Cortex* papers to about 68 references in the most cited *Brain* papers.

Fourth, there is no clear overall trend in citations as a function of institutional representation. The institution/ (institution + university) ratio (where institution in the table cells should be interpreted as any non-university organization; e.g., research laboratory, clinic, hospital, company) for most cited papers starts at 0.5 for Cortex, drops to 0.2 for *Neuropsychologia*, and increases sharply to 0.8 for *Brain*. This ratio for least cited papers starts at 0.4 for both *Cortex* and *Neuropsychologia*, and decreases to 0.2 for *Brain*. Its most dramatic change is from 0.8 for the most cited *Brain* papers to 0.2 for the least cited *Brain* papers.

Fifth, the most cited papers in *Cortex* are all from continental Western Europe, with heavy representation from Italy and France, while the least cited papers in *Cortex* represent four different continents. The most cited papers in *Neuropsychologia* are, with the exception of Italy, from the UK and North America (with heavy representation from the UK and USA), while the least cited papers have more representation from Western Europe but none from the UK. The most cited papers in *Brain* are from the major English-speaking countries, whereas the least cited are scattered around Western Europe, Asia, and North America.

Sixth, there is a distinct shift in type of study (the bottom of Table 7) in proceeding from *Cortex* to *Neuropsychologia* to *Brain*. Clinical behavioral studies, many of them essentially case studies, predominate the most cited Cortex papers. There are only two papers characterized as Diagnostic-Non-Invasive (e.g., PET, MRI, etc). *Neuropsychologia* has more of a balance between Behavioral and Diagnostic-Non-Invasive in its ten most cited papers. *Brain* shows a heavy emphasis on Diagnostic-Non-Invasive (7/10), two papers on surgical procedures, and one on Diagnostic-Invasive. Based on reading Abstracts from each of these journals, the types as represented in the top ten most cited articles roughly approximate the types of papers published overall. Thus, as citations increase in absolute amounts, the study type transitions from the clinically oriented behavioral focus to the correlates with more objective measurements. Also, as the results from the most cited papers section showed, as the study type transitions from the clinically oriented behavioral focus ('soft' technology) to the more objective measurements ('hard' technology), the most cited papers tend to become more recent.

Section 14. Biomedical Text Mining Bibliography

Based on: analyses performed in late 2005; Kostoff, R.N.)

This section contains a bibliography of articles emphasizing text mining. Because the fundamental principles of biomedical text mining are similar to those of science and technology text mining, the articles in the bibliography are not constrained to biomedical text mining, but encompass all of text mining. While the bibliography contains some articles on Information Retrieval and Natural Language Processing, it does not cover the more general articles in these literatures.

Before the 2005 and prior article citations are presented, some overall bibliometric quantities will be presented, to provide an overview of the text mining field's infrastructure. To obtain the bibliometrics, the text mining articles retrieved were subjected to bibliometric and phrase frequency analyses. These bibliometric quantities are presented in frequency order, highest first. In the first category presented (Important Phrases), the ordering is by number of words in phrase, then by frequency.

QUERY USED TO RETRIEVE 2005 AND PRIOR TEXT MINING ARTICLES

(search engine* OR relevance feedback OR document retrieval OR information filtering OR database tomography OR information retrieval OR (query SAME (TREC OR web OR retrieval performance OR relevant document* OR vector space model* OR document collection* OR term independence OR syntactic term mismatch* OR coordinated index concept* OR text OR texts OR (n gram* OR ngram*) OR information OR document* OR language OR boolean OR LSI)) OR ((text OR texts) SAME (search* OR knowledge discovery OR retrieval OR queries OR information extraction OR mining)) OR (search* SAME (TREC OR bitext OR textual OR terms OR hypertext OR (fulltext OR full text OR freetext OR free text) OR hotbot OR LSI OR YAHOO OR relevance judgment* OR ALTAVISTA OR bibliographic OR boolean OR internet OR world wide web OR natural language OR keyword*)) OR (data mining SAME (document* OR term frequenc*)) OR (text mining) OR (document* SAME (LSI OR queries OR querying)) OR (documents SAME (search* OR boolean OR clustering) OR ((free text OR freetext OR full text OR fulltext) SAME retrieval) OR (latent semantic indexing SAME (search* OR retrieval OR document*)) OR (queries SAME (textual OR internet OR TREC)) OR ((searches OR

searching) SAME (documents OR TREC)))) NOT (gene OR forest OR fuzzy sets OR fuzzy rules OR fuzzy systems OR fuzzy logic OR frames OR heuristics OR logic program* OR scheduling OR detection OR data set* OR signals OR spatial data OR simulations OR (rulebased OR rule based) OR rule induction OR time series OR temperature OR speech recognition OR transactions OR optical OR numerical OR multivariate OR molecular OR membership functions OR regression OR radio OR pruning OR polynomial OR planning OR petri OR itemset* OR configuration OR horn OR hiv OR audio OR atomic OR expert system* OR KBS OR operator OR (search SAME (evidence OR clinical OR sequence* OR video OR trial* OR treatment* OR patients OR protein* OR object* OR temporal OR folate OR drug* OR disease* OR depression OR data synthesis OR mortality OR injury OR morbidity OR medications OR image* OR hypertension OR cancer OR blood OR blast OR breast OR atrial OR acid OR caries OR chemical OR risk OR therapy OR motion OR tabu OR stroke OR pressure OR program OR multimedia OR randomized)) OR (rules SAME (database OR query)) OR (query SAME (object* OR linear OR pictorial OR algebra* OR amino OR DNA OR genome OR graphical OR temporal OR signature* OR video data OR visual OR sequences OR program OR protein*)) OR (queries SAME (linear OR algorithm*)) OR (program SAME (database OR sequence* OR protein*)) OR (network* SAME (model* OR video OR distributed)) OR (knowledge SAME objects) OR (image* SAME (information OR video OR data OR algorithm* OR feature* OR matching OR model OR content OR databases OR digital OR motion OR shape OR multimedia OR representation OR object* OR processing)) OR (fuzzy SAME data) OR (data SAME (folate OR depression OR disease* OR factors OR complexity OR genome OR glucose OR biological OR infection* OR acid OR articles OR costs OR attributes OR blood OR children OR surgery OR sexual OR shape OR therapy OR transmission OR warfarin OR calcium OR medline OR stroke OR pulmonary OR morbidity OR mortality OR neural OR learning OR pregnancy OR mutations OR object OR objects OR drug* OR human OR treatment* OR trial* OR patients)))

UPDATE FOR DATA OBTAINED IN OCTOBER 2007
(ABBREVIATED QUERY USED)

PROLIFIC TEXT MINING AUTHORS

Author	Record Count	% of 895
KOSTOFF, RN	13	1.4525%

CHEN, HC	7	0.7821%
KATOH, M	7	0.7821%
CHAU, M	6	0.6704%
DEGEMMIS, M	6	0.6704%
LOPS, P	6	0.6704%
BREMER, EG	5	0.5587%
CHIEN, LF	5	0.5587%
DESESA, C	5	0.5587%
FUKETA, M	5	0.5587%
HIRSCHMAN, L	5	0.5587%
KARYPIS, G	5	0.5587%
LEE, GG	5	0.5587%
LI, S	5	0.5587%
NATARAJAN, J	5	0.5587%
SEMERARO, G	5	0.5587%
TAN, SB	5	0.5587%
VALENCIA, A	5	0.5587%
ADEVA, JJG	4	0.4469%
ATLAM, ES	4	0.4469%
BLASCHKE, C	4	0.4469%
BORK, P	4	0.4469%
CHEN, H	4	0.4469%
CHEN, X	4	0.4469%
CHIANG, JH	4	0.4469%
CORCHADO, JM	4	0.4469%
DIAZ, F	4	0.4469%
DUBITZKY, W	4	0.4469%
FDEZ-RIVEROLA, F	4	0.4469%
HACK, CJ	4	0.4469%
HU, XH	4	0.4469%
HU, ZZ	4	0.4469%
IGLESIAS, EL	4	0.4469%
KAMEL, MS	4	0.4469%
KIMURA, F	4	0.4469%
KRALLINGER, M	4	0.4469%
LEE, CH	4	0.4469%
LI, JZ	4	0.4469%
LIU, Y	4	0.4469%
MA, WY	4	0.4469%
MALPOHL, G	4	0.4469%

MENDEZ, JR 4 0.4469%
MONS, B 4 0.4469%
MORITA, K 4 0.4469%
OHYAMA, W 4 0.4469%
PAL, SK 4 0.4469%
RYU, KH 4 0.4469%
SCHNEIDER, KM 4 0.4469%
SMALHEISER, NR 4 0.4469%
SRINIVASAN, P 4 0.4469%
TANG, J 4 0.4469%
TSHITEYA, R 4 0.4469%
WAKABAYASHI, T 4 0.4469%
WANG, C 4 0.4469%
WANG, DL 4 0.4469%
WEI, CP 4 0.4469%
WU, CH 4 0.4469%
XU, H 4 0.4469%
XU, J 4 0.4469%
YOON, Y 4 0.4469%
YU, Y 4 0.4469%
ZHANG, J 4 0.4469%
ZHOU, ZH 4 0.4469%
ABULAISH, M 3 0.3352%
ANANIADOU, S 3 0.3352%
AOE, JI 3 0.3352%
BAO, YB 3 0.3352%
BI, YX 3 0.3352%
CARRASCO-OCHOA, JA 3 0.3352%
CHEN, J 3 0.3352%
CHEN, SM 3 0.3352%
CHEN, WL 3 0.3352%
CHEN, Y 3 0.3352%
CHIANG, IJ 3 0.3352%
COLOSIMO, M 3 0.3352%
COMBARRO, EF 3 0.3352%
CORREA, RF 3 0.3352%
CORTES, HD 3 0.3352%
DEL RIO, JA 3 0.3352%
DENG, ZH 3 0.3352%
DEY, L 3 0.3352%

DIAZ, I 3 0.3352%
FELDMAN, R 3 0.3352%
FREEMAN, RT 3 0.3352%
FRIEDMAN, C 3 0.3352%
GAO, W 3 0.3352%
GARCIA-HERNANDEZ, RA 3 0.3352%
HUANG, HK 3 0.3352%
HWANG, JH 3 0.3352%
JENSEN, LJ 3 0.3352%
KATOH, Y 3 0.3352%
KIM, HJ 3 0.3352%
KIM, S 3 0.3352%
KLOPOTEK, MA 3 0.3352%
KORS, JA 3 0.3352%
KOYTCHEFF, RG 3 0.3352%
LAM, W 3 0.3352%
LAU, CGY 3 0.3352%
LEE, C 3 0.3352%
LEE, JH 3 0.3352%
LEE, YS 3 0.3352%
LEWIS, DD 3 0.3352%
LI, Q 3 0.3352%
LI, XG 3 0.3352%
LI, YF 3 0.3352%
LIN, YM 3 0.3352%
LIU, HF 3 0.3352%
LUDERMIR, TB 3 0.3352%
MARTIN-MERINO, M 3 0.3352%
MARTINEZ-TRINIDAD, JF 3 0.3352%
MONTANES, E 3 0.3352%
MORGAN, AA 3 0.3352%
MYAENG, SH 3 0.3352%
NARAYANASWAMY, M 3 0.3352%
PARK, J 3 0.3352%
QIAN, TY 3 0.3352%
QU, YL 3 0.3352%
RANILLA, J 3 0.3352%
RAVIKUMAR, KE 3 0.3352%
RZHETSKY, A 3 0.3352%
SEBASTIANI, F 3 0.3352%

SHANG, WQ	3	0.3352%
SHATKAY, H	3	0.3352%
SILVA, MJ	3	0.3352%
TSUJII, J	3	0.3352%
VIDAL, E	3	0.3352%
VIJAY-SHANKER, K	3	0.3352%
WANG, H	3	0.3352%
WANG, SY	3	0.3352%
WANG, YZ	3	0.3352%
YEH, A	3	0.3352%
YU, G	3	0.3352%
ZHANG, M	3	0.3352%
ZHU, HB	3	0.3352%
ZHU, JB	3	0.3352%
ZU, GW	3	0.3352%
ZUO, WL	3	0.3352%
ABE, K	2	0.2235%
ABRAHAM, A	2	0.2235%
ALIFERIS, CF	2	0.2235%
ALMONAYYES, A	2	0.2235%
AMANDI, A	2	0.2235%
AOE, J	2	0.2235%
APHINYANAPHONGS, Y	2	0.2235%
ATXA, JMP	2	0.2235%
BAJIC, VB	2	0.2235%
BAKER, CJO	2	0.2235%
BAO, YG	2	0.2235%
BARAL, C	2	0.2235%
BASILE, TMA	2	0.2235%
BASILI, R	2	0.2235%
BEITZEL, SM	2	0.2235%
BELL, D	2	0.2235%
BERENDT, B	2	0.2235%
BERGER, H	2	0.2235%
BERRAR, D	2	0.2235%
BHIMAVARAPU, R	2	0.2235%
BICHINDARITZ, I	2	0.2235%
BITTNER, M	2	0.2235%
BOBERG, J	2	0.2235%
BORZEMSKI, L	2	0.2235%

BRATKO, A	2	0.2235%
CALERA-RUBIO, J	2	0.2235%
CARDOSO, N	2	0.2235%
CHANG, HC	2	0.2235%
CHAVES, M	2	0.2235%
CHEN, DY	2	0.2235%
CHEN, EH	2	0.2235%
CHEN, XY	2	0.2235%
CHEN, Z	2	0.2235%
CHENG, T	2	0.2235%
CHENG, TH	2	0.2235%
CHOU, WC	2	0.2235%
CHOWDHURY, A	2	0.2235%
CHUA, TS	2	0.2235%
CHUANG, SL	2	0.2235%
CIESIELSKI, K	2	0.2235%
CIVERA, J	2	0.2235%
COHEN, AM	2	0.2235%
COLLIER, N	2	0.2235%
COLOMBE, JB	2	0.2235%
CORMACK, GV	2	0.2235%
CUI, XH	2	0.2235%
CZERSKI, D	2	0.2235%
DAI, HH	2	0.2235%
DALAMAGAS, T	2	0.2235%
DAVULCU, H	2	0.2235%
DE MOOR, B	2	0.2235%
DEBENHAM, J	2	0.2235%
DEBNATH, S	2	0.2235%
DI MAURO, N	2	0.2235%
DOAN, S	2	0.2235%
DOBRYNIN, V	2	0.2235%
DONG, ZY	2	0.2235%
DRAMINSKI, M	2	0.2235%
ERHARDT, RAA	2	0.2235%
ESULI, A	2	0.2235%
FAGNI, T	2	0.2235%
FAN, WG	2	0.2235%
FENG, BQ	2	0.2235%
FENG, JL	2	0.2235%

FERILLI, S	2	0.2235%
FILIPIC, B	2	0.2235%
FLESCA, S	2	0.2235%
FRANK, E	2	0.2235%
FRESKO, M	2	0.2235%
FRIEDER, O	2	0.2235%
GALUSHKA, M	2	0.2235%
GHOSH, J	2	0.2235%
GILES, CL	2	0.2235%
GINTER, F	2	0.2235%
GODOY, D	2	0.2235%
GORDON, MD	2	0.2235%
GUAN, JW	2	0.2235%
GUO, GD	2	0.2235%
GUO, J	2	0.2235%
GUO, L	2	0.2235%
GUO, Y	2	0.2235%
HAFFNER, P	2	0.2235%
HAN, H	2	0.2235%
HE, QM	2	0.2235%
HE, YL	2	0.2235%
HERNANDEZ-REYES, E	2	0.2235%
HERSH, WR	2	0.2235%
HIDALGO, JMG	2	0.2235%
HIRSCH, L	2	0.2235%
HIRSCH, R	2	0.2235%
HONAVAR, V	2	0.2235%
HONG, K	2	0.2235%
HORIGUCHI, S	2	0.2235%
HOTHO, A	2	0.2235%
HSIANG, J	2	0.2235%
HSU, CC	2	0.2235%
HSU, WL	2	0.2235%
HU, HP	2	0.2235%
HU, JS	2	0.2235%
HU, YF	2	0.2235%
HUANG, J	2	0.2235%
HUANG, S	2	0.2235%
HUANG, XC	2	0.2235%
HUI, SC	2	0.2235%

HUNG, CM	2	0.2235%
INESTA, JM	2	0.2235%
IOSSIFOV, I	2	0.2235%
ISHII, N	2	0.2235%
IVANOVIC, M	2	0.2235%
JARVINEN, J	2	0.2235%
JENSEN, EC	2	0.2235%
JIANG, MH	2	0.2235%
JIN, Y	2	0.2235%
JONES, R	2	0.2235%
JOO, KH	2	0.2235%
JUAN, A	2	0.2235%
KADOYA, Y	2	0.2235%
KAN, MY	2	0.2235%
KANDEL, A	2	0.2235%
KANG, BY	2	0.2235%
KANG, DK	2	0.2235%
KASHIJI, S	2	0.2235%
KELL, DB	2	0.2235%
KIM, SB	2	0.2235%
KIRSCH, H	2	0.2235%
KOHLER, J	2	0.2235%
KOPPEL, M	2	0.2235%
KUFLIK, T	2	0.2235%
LAI, KK	2	0.2235%
LAST, M	2	0.2235%
LEE, D	2	0.2235%
LEE, KH	2	0.2235%
LEE, M	2	0.2235%
LEE, WS	2	0.2235%
LERTNATTEE, V	2	0.2235%
LEVENE, M	2	0.2235%
LEYDESDORFF, L	2	0.2235%
LI, CH	2	0.2235%
LI, QZ	2	0.2235%
LI, RL	2	0.2235%
LI, T	2	0.2235%
LI, X	2	0.2235%
LI, XX	2	0.2235%
LI, ZH	2	0.2235%

LIANG, BY	2	0.2235%
LIANG, CY	2	0.2235%
LIAO, SS	2	0.2235%
LIN, TY	2	0.2235%
LISI, FA	2	0.2235%
LIU, B	2	0.2235%
LIU, H	2	0.2235%
LIU, J	2	0.2235%
LIU, RL	2	0.2235%
LIU, SZ	2	0.2235%
LIU, T	2	0.2235%
LIU, TY	2	0.2235%
LOH, HT	2	0.2235%
LU, W	2	0.2235%
LU, YC	2	0.2235%
LUO, X	2	0.2235%
LYNAM, TR	2	0.2235%
MA, ZF	2	0.2235%
MALIK, R	2	0.2235%
MARCELLONI, F	2	0.2235%
MARCOTTE, EM	2	0.2235%
MARKATOU, M	2	0.2235%
MARTIN, J	2	0.2235%
MARTINS, B	2	0.2235%
MATSUMOTO, Y	2	0.2235%
MATSUO, Y	2	0.2235%
MATWIN, S	2	0.2235%
MENCZER, F	2	0.2235%
MENG, B	2	0.2235%
MILANESI, L	2	0.2235%
MONTES-Y-GOMEZ, M	2	0.2235%
MOONEY, RJ	2	0.2235%
MORGAN, A	2	0.2235%
MOSCHITTI, A	2	0.2235%
MUNOZ, A	2	0.2235%
MURAI, T	2	0.2235%
MURATA, M	2	0.2235%
NAM, T	2	0.2235%
NAPOLI, A	2	0.2235%
NENADIC, G	2	0.2235%

NG, MK	2	0.2235%
NIE, JY	2	0.2235%
OGIHARA, M	2	0.2235%
OKUMURA, M	2	0.2235%
PALAKAL, M	2	0.2235%
PAN, H	2	0.2235%
PANT, G	2	0.2235%
PARK, JC	2	0.2235%
PARK, SC	2	0.2235%
PATHAK, P	2	0.2235%
PATTERSON, D	2	0.2235%
PENG, H	2	0.2235%
PEREZ-SANCHO, C	2	0.2235%
PFAHRINGER, B	2	0.2235%
PHILIPPI, S	2	0.2235%
PONS-PORRATA, A	2	0.2235%
POTOK, TE	2	0.2235%
QIN, JL	2	0.2235%
QUARESMA, P	2	0.2235%
RADOVANOVIC, M	2	0.2235%
RAMAKRISHNAN, IV	2	0.2235%
RAMANI, AK	2	0.2235%
REBHOLZ-SCHUHMANN, D	2	0.2235%
RIM, HC	2	0.2235%
ROONEY, N	2	0.2235%
ROSENFELD, B	2	0.2235%
ROUSU, J	2	0.2235%
RUCH, P	2	0.2235%
RUEGG, A	2	0.2235%
SAEEDI, M	2	0.2235%
SALAKOSKI, T	2	0.2235%
SCHUEMIE, MJ	2	0.2235%
SELLIS, T	2	0.2235%
SHAH, PK	2	0.2235%
SHAWE-TAYLOR, J	2	0.2235%
SHEN, H	2	0.2235%
SHIN, K	2	0.2235%
SIEBES, A	2	0.2235%
SILVA, C	2	0.2235%
SILVESCU, A	2	0.2235%

SINGH, P	2	0.2235%
SMITH, A	2	0.2235%
SMITH, C	2	0.2235%
SONG, DW	2	0.2235%
SONG, IY	2	0.2235%
SPECHT, M	2	0.2235%
STAMATATOS, E	2	0.2235%
STANIMIROVIC, DB	2	0.2235%
STATNIKOV, A	2	0.2235%
SUMITOMO, T	2	0.2235%
SUNG, TY	2	0.2235%
SWANSON, DR	2	0.2235%
TAKAGI, T	2	0.2235%
TAKAHASHI, H	2	0.2235%
TAKAHASHI, M	2	0.2235%
TAKAHASHI, S	2	0.2235%
TAKAMURA, H	2	0.2235%
TAKEUCHI, K	2	0.2235%
TAN, HZ	2	0.2235%
TANAKA, K	2	0.2235%
TANG, N	2	0.2235%
TANG, SW	2	0.2235%
TEZUKA, T	2	0.2235%
THEERAMUNKONG, T	2	0.2235%
TORII, M	2	0.2235%
TORVIK, VI	2	0.2235%
TOUSSAINT, Y	2	0.2235%
TSUDA, K	2	0.2235%
VELASQUEZ, JD	2	0.2235%
VILLASENOR-PINEDA, L	2	0.2235%
WAGNER, C	2	0.2235%
WANG, JB	2	0.2235%
WANG, JY	2	0.2235%
WANG, K	2	0.2235%
WANG, KH	2	0.2235%
WANG, T	2	0.2235%
WANG, W	2	0.2235%
WANG, ZH	2	0.2235%
WEEBER, M	2	0.2235%
WEIKUM, G	2	0.2235%

WERNER, T	2	0.2235%
WHITE, PS	2	0.2235%
WIERZCHON, ST	2	0.2235%
WILBUR, WJ	2	0.2235%
WINKEL, KJ	2	0.2235%
WINTERS, RS	2	0.2235%
WITTE, R	2	0.2235%
WREN, JD	2	0.2235%
WU, LJ	2	0.2235%
WU, SH	2	0.2235%
WU, ZH	2	0.2235%
XU, C	2	0.2235%
XU, P	2	0.2235%
XUE, GR	2	0.2235%
YAMADA, T	2	0.2235%
YAN, PL	2	0.2235%
YANG, HC	2	0.2235%
YANG, J	2	0.2235%
YANG, ZY	2	0.2235%
YAO, TS	2	0.2235%
YEH, AS	2	0.2235%
YIN, HJ	2	0.2235%
YOO, I	2	0.2235%
YU, H	2	0.2235%
YU, L	2	0.2235%
YU, PS	2	0.2235%
YU, S	2	0.2235%
ZAKI, MJ	2	0.2235%
ZELIKOVITZ, S	2	0.2235%
ZHANG, ML	2	0.2235%
ZHANG, Q	2	0.2235%
ZHANG, YH	2	0.2235%
ZHAO, Y	2	0.2235%
ZHONG, N	2	0.2235%
ZHONG, S	2	0.2235%
ZHOU, XZ	2	0.2235%
ZHOU, YL	2	0.2235%
ZHU, QS	2	0.2235%
ZHUANG, L	2	0.2235%

PROLIFIC TEXT MINING COUNTRIES

Country/Territory	Record Count	% of 895
USA	227	25.3631%
PEOPLES R CHINA	149	16.6480%
JAPAN	74	8.2682%
GERMANY	56	6.2570%
SOUTH KOREA	54	6.0335%
ENGLAND	49	5.4749%
SPAIN	43	4.8045%
TAIWAN	43	4.8045%
CANADA	39	4.3575%
ITALY	35	3.9106%
AUSTRALIA	29	3.2402%
SINGAPORE	22	2.4581%
INDIA	20	2.2346%
FRANCE	18	2.0112%
ISRAEL	18	2.0112%
NETHERLANDS	18	2.0112%
GREECE	17	1.8994%
POLAND	15	1.6760%
SWITZERLAND	12	1.3408%
NORTH IRELAND	11	1.2291%
PORTUGAL	10	1.1173%
MEXICO	9	1.0056%
BRAZIL	8	0.8939%
FINLAND	7	0.7821%
TURKEY	7	0.7821%
THAILAND	6	0.6704%
NEW ZEALAND	5	0.5587%
SLOVENIA	5	0.5587%
BELGIUM	4	0.4469%
AUSTRIA	3	0.3352%
CHILE	3	0.3352%
CZECH REPUBLIC	3	0.3352%
HUNGARY	3	0.3352%
IRELAND	3	0.3352%
NORWAY	3	0.3352%
RUSSIA	3	0.3352%
SWEDEN	3	0.3352%

ARGENTINA	2	0.2235%
CUBA	2	0.2235%
EGYPT	2	0.2235%
JORDAN	2	0.2235%
KUWAIT	2	0.2235%
SCOTLAND	2	0.2235%
SERBIA	2	0.2235%
U ARAB EMIRATES	2	0.2235%
VIETNAM	2	0.2235%
IRAN	1	0.1117%
LITHUANIA	1	0.1117%
MALAYSIA	1	0.1117%
OMAN	1	0.1117%
QATAR	1	0.1117%
SLOVAKIA	1	0.1117%
SOUTH AFRICA	1	0.1117%

PROLIFIC TEXT MINING INSTITUTIONS

Institution Name	Record Count	% of 895
TSING HUA UNIV	21	2.3464%
CHINESE ACAD SCI	17	1.8994%
OFF NAVAL RES	13	1.4525%
UNIV TOKYO	11	1.2291%
UNIV BARI	10	1.1173%
UNIV ILLINOIS	10	1.1173%
UNIV ULSTER	10	1.1173%
ACAD SINICA	9	1.0056%
NANYANG TECHNOL UNIV	9	1.0056%
NATL UNIV SINGAPORE	9	1.0056%
UNIV ARIZONA	9	1.0056%
CARNEGIE MELLON UNIV	8	0.8939%
COLUMBIA UNIV	8	0.8939%
UNIV MINNESOTA	8	0.8939%
UNIV QUEENSLAND	8	0.8939%
UNIV WATERLOO	8	0.8939%
BEN GURION UNIV NEGEV	7	0.7821%
CNR	7	0.7821%
CSIC	7	0.7821%
KOREA ADV INST SCI & TECHNOL	7	0.7821%

NATL CHENG KUNG UNIV	7	0.7821%
NORTHEASTERN UNIV	7	0.7821%
PEKING UNIV	7	0.7821%
POLISH ACAD SCI	7	0.7821%
UNIV HONG KONG	7	0.7821%
UNIV KARLSRUHE	7	0.7821%
UNIV MANCHESTER	7	0.7821%
UNIV SALAMANCA	7	0.7821%
ZHEJIANG UNIV	7	0.7821%
BAR ILAN UNIV	6	0.6704%
CHINESE UNIV HONG KONG	6	0.6704%
DREXEL UNIV	6	0.6704%
FUDAN UNIV	6	0.6704%
HUAZHONG UNIV SCI & TECHNOL	6	0.6704%
INDIAN INST TECHNOL	6	0.6704%
INST INFOCOMM RES	6	0.6704%
M&M MED BIOINFORMAT	6	0.6704%
NATL CANC CTR	6	0.6704%
NATL TAIWAN UNIV	6	0.6704%
NATL TAIWAN UNIV SCI & TECHNOL	6	0.6704%
POHANG UNIV SCI & TECHNOL	6	0.6704%
SHANGHAI JIAO TONG UNIV	6	0.6704%
UNIV CALIF DAVIS	6	0.6704%
WUHAN UNIV	6	0.6704%
CHONBUK NATL UNIV	5	0.5587%
EUROPEAN MOL BIOL LAB	5	0.5587%
GEORGETOWN UNIV	5	0.5587%
HARBIN INST TECHNOL	5	0.5587%
HONG KONG POLYTECH UNIV	5	0.5587%
IBM CORP	5	0.5587%
INDIANA UNIV	5	0.5587%
KOREA UNIV	5	0.5587%
MITRE CORP	5	0.5587%
NATL RES COUNCIL CANADA	5	0.5587%
PENN STATE UNIV	5	0.5587%
RUTGERS STATE UNIV	5	0.5587%
UNIV IOWA	5	0.5587%
UNIV SYDNEY	5	0.5587%
UNIV TEXAS	5	0.5587%
UNIV TOKUSHIMA	5	0.5587%

UNIV VIGO 5 0.5587%
UNIV WASHINGTON 5 0.5587%
WROCLAW TECH UNIV 5 0.5587%
ANNA UNIV 4 0.4469%
BEIJING UNIV POSTS & TELECOMMUN 4 0.4469%
CHUNGBUK NATL UNIV 4 0.4469%
DDL OMNI ENGN LLC 4 0.4469%
FLORIDA INT UNIV 4 0.4469%
GEORGIA INST TECHNOL 4 0.4469%
HARVARD UNIV 4 0.4469%
HEBREW UNIV JERUSALEM 4 0.4469%
IIT 4 0.4469%
INDIAN STAT INST 4 0.4469%
INFORMAT & COMMUN UNIV 4 0.4469%
JILIN UNIV 4 0.4469%
JOZEF STEFAN INST 4 0.4469%
MAX DELBRUCK CTR MOL MED 4 0.4469%
MICROSOFT RES ASIA 4 0.4469%
MIE UNIV 4 0.4469%
MIT 4 0.4469%
NANJING UNIV 4 0.4469%
NATL TECH UNIV ATHENS 4 0.4469%
NATL TSING HUA UNIV 4 0.4469%
NEW JERSEY INST TECHNOL 4 0.4469%
NORTHWESTERN UNIV 4 0.4469%
OSAKA UNIV 4 0.4469%
S CHINA UNIV TECHNOL 4 0.4469%
UNIV AMSTERDAM 4 0.4469%
UNIV BASQUE COUNTRY 4 0.4469%
UNIV BIELEFELD 4 0.4469%
UNIV DELAWARE 4 0.4469%
UNIV MICHIGAN 4 0.4469%
UNIV OTTAWA 4 0.4469%
UNIV PASSAU 4 0.4469%
UNIV PENN 4 0.4469%
UNIV POLITECN VALENCIA 4 0.4469%
UNIV TECHNOL SYDNEY 4 0.4469%
UNIV TSUKUBA 4 0.4469%
UNIV VALLADOLID 4 0.4469%
XIAN JIAOTONG UNIV 4 0.4469%

BEIJING JIAOTONG UNIV 3 0.3352%
BIOALMA 3 0.3352%
CHAOYANG UNIV TECHNOL 3 0.3352%
CITY UNIV HONG KONG 3 0.3352%
CNRS 3 0.3352%
CONCORDIA UNIV 3 0.3352%
ELECT & TELECOMMUN RES INST 3 0.3352%
FUZHOU UNIV 3 0.3352%
HARBIN ENGN UNIV 3 0.3352%
HUMBOLDT UNIV 3 0.3352%
INAOE 3 0.3352%
IOWA STATE UNIV 3 0.3352%
KATHOLIEKE UNIV LEUVEN 3 0.3352%
KYOTO UNIV 3 0.3352%
LEIDEN UNIV 3 0.3352%
MONASH UNIV 3 0.3352%
NARA INST SCI & TECHNOL 3 0.3352%
NATL KAOHSIUNG UNIV APPL SCI 3 0.3352%
NATL SUN YAT SEN UNIV 3 0.3352%
OKAYAMA UNIV 3 0.3352%
OREGON HLTH & SCI UNIV 3 0.3352%
QUEENS UNIV 3 0.3352%
QUEENS UNIV BELFAST 3 0.3352%
SEOUL NATL UNIV 3 0.3352%
SOGANG UNIV 3 0.3352%
STANFORD UNIV 3 0.3352%
TAIPEI MED UNIV 3 0.3352%
TIANJIN UNIV 3 0.3352%
UNAM 3 0.3352%
UNIV CALIF BERKELEY 3 0.3352%
UNIV CHICAGO 3 0.3352%
UNIV COLL LONDON 3 0.3352%
UNIV EUROPEA MADRID 3 0.3352%
UNIV FED PERNAMBUCO 3 0.3352%
UNIV FLORIDA 3 0.3352%
UNIV HELSINKI 3 0.3352%
UNIV LISBON 3 0.3352%
UNIV MARYLAND 3 0.3352%
UNIV MASSACHUSETTS 3 0.3352%
UNIV NEW S WALES 3 0.3352%

UNIV OVIEDO 3 0.3352%
UNIV PITTSBURGH 3 0.3352%
UNIV SCI & TECHNOL CHINA 3 0.3352%
UNIV SEOUL 3 0.3352%
UNIV SO CALIF 3 0.3352%
UNIV SOUTHAMPTON 3 0.3352%
UNIV TENNESSEE 3 0.3352%
UNIV TWENTE 3 0.3352%
UNIV UTAH 3 0.3352%
UNIV WAIKATO 3 0.3352%
UNIV ZURICH 3 0.3352%
VIRGINIA POLYTECH INST & STATE UNIV 3 0.3352%
AICHI INST TECHNOL 2 0.2235%
AIST 2 0.2235%
ARISTOTLE UNIV THESSALONIKI 2 0.2235%
ARIZONA STATE UNIV 2 0.2235%
AT&T LABS RES 2 0.2235%
ATHENS UNIV ECON & BUSINESS 2 0.2235%
BENTLEY COLL 2 0.2235%
BIOALMA SL 2 0.2235%
BOOZ ALLEN HAMILTON 2 0.2235%
CHANG JUNG UNIV 2 0.2235%
CHILDRENS HOSP PHILADELPHIA 2 0.2235%
CHONGQING UNIV 2 0.2235%
CHUNG ANG UNIV 2 0.2235%
CUNY COLL STATEN ISL 2 0.2235%
DALHOUSIE UNIV 2 0.2235%
DAVID D LEWIS CONSULTING 2 0.2235%
DEAKIN UNIV 2 0.2235%
DONG A UNIV 2 0.2235%
ERASMUS UNIV 2 0.2235%
FLORIDA ATLANTIC UNIV 2 0.2235%
FUJIAN NORMAL UNIV 2 0.2235%
GYEONGIN NATL UNIV EDUC 2 0.2235%
HOKKAIDO UNIV 2 0.2235%
HUNAN UNIV 2 0.2235%
HUNGARIAN ACAD SCI 2 0.2235%
IDIAP RES INST 2 0.2235%
INDIANA UNIV PURDUE UNIV 2 0.2235%
INSERM 2 0.2235%

INST DEF ANAL	2	0.2235%
JADAVPUR UNIV	2	0.2235%
JAMIA MILLIA ISLAMIA	2	0.2235%
JAPAN ADV INST SCI & TECHNOL	2	0.2235%
KUWAIT UNIV	2	0.2235%
MAEBASHI INST TECHNOL	2	0.2235%
MING CHUAN UNIV	2	0.2235%
MITSUI ASSET TRUST & BANKING CO LTD	2	0.2235%
NATL CENT UNIV	2	0.2235%
NATL CHIAO TUNG UNIV	2	0.2235%
NATL CTR TEXT MIN	2	0.2235%
NATL INST INFORMAT	2	0.2235%
NATL LIB MED	2	0.2235%
NATL TAIWAN NORMAL UNIV	2	0.2235%
NATL UNIV DEF TECHNOL	2	0.2235%
NIPISSING UNIV	2	0.2235%
NTT CORP	2	0.2235%
OAK RIDGE NATL LAB	2	0.2235%
OPEN UNIV	2	0.2235%
POLITECN MILAN	2	0.2235%
POZNAN TECH UNIV	2	0.2235%
QUEENSLAND UNIV TECHNOL	2	0.2235%
RADBOUD UNIV NIJMEGEN	2	0.2235%
RENSSELAER POLYTECH INST	2	0.2235%
SAN JOSE STATE UNIV	2	0.2235%
SELCUK UNIV	2	0.2235%
SHIMANE UNIV	2	0.2235%
SILPAKORN UNIV	2	0.2235%
SIMON FRASER UNIV	2	0.2235%
SINGAPORE MANAGEMENT UNIV	2	0.2235%
ST PETERSBURG STATE UNIV	2	0.2235%
SWISS FED INST TECHNOL	2	0.2235%
TEL AVIV UNIV	2	0.2235%
TOHOKU UNIV	2	0.2235%
TOKYO INST TECHNOL	2	0.2235%
TOSHIBA SOLUT CORP	2	0.2235%
TRANSLAT GENOM RES INST	2	0.2235%
UMIST	2	0.2235%
UNIV AEGEAN	2	0.2235%
UNIV ALBERTA	2	0.2235%

UNIV ALICANTE 2 0.2235%
UNIV CALABRIA 2 0.2235%
UNIV CALIF LOS ANGELES 2 0.2235%
UNIV CALIF SANTA BARBARA 2 0.2235%
UNIV CARLOS III MADRID 2 0.2235%
UNIV CINCINNATI 2 0.2235%
UNIV DORTMUND 2 0.2235%
UNIV DUBLIN TRINITY COLL 2 0.2235%
UNIV ELECT SCI & TECHNOL CHINA 2 0.2235%
UNIV EVORA 2 0.2235%
UNIV HOSP GENEVA 2 0.2235%
UNIV JAEN 2 0.2235%
UNIV JAUME 1 2 0.2235%
UNIV KASSEL 2 0.2235%
UNIV LJUBLJANA 2 0.2235%
UNIV LONDON 2 0.2235%
UNIV LONDON ROYAL HOLLOWAY & BEDFORD NEW COLL 2 0.2235%
UNIV MED CTR ROTTERDAM 2 0.2235%
UNIV MILAN 2 0.2235%
UNIV MISSOURI 2 0.2235%
UNIV MONTREAL 2 0.2235%
UNIV MUNICH 2 0.2235%
UNIV NEBRASKA 2 0.2235%
UNIV NOVI SAD 2 0.2235%
UNIV OKLAHOMA 2 0.2235%
UNIV ORIENTE 2 0.2235%
UNIV PADUA 2 0.2235%
UNIV PATRAS 2 0.2235%
UNIV PISA 2 0.2235%
UNIV PORTO 2 0.2235%
UNIV ROCHESTER 2 0.2235%
UNIV ROMA TOR VERGATA 2 0.2235%
UNIV S FLORIDA 2 0.2235%
UNIV SAO PAULO 2 0.2235%
UNIV SHEFFIELD 2 0.2235%
UNIV SURREY 2 0.2235%
UNIV TURKU 2 0.2235%
UNIV UTRECHT 2 0.2235%
UNIV WISCONSIN 2 0.2235%

VANDERBILT UNIV 2 0.2235%
YALE UNIV 2 0.2235%
YONSEI UNIV 2 0.2235%

PROLIFIC TEXT MINING JOURNALS

Source Title Record Count % of 895
BMC BIOINFORMATICS 38 4.2458%
INFORMATION PROCESSING & MANAGEMENT 30 3.3520%
IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING 28 3.1285%
BIOINFORMATICS 19 2.1229%
EXPERT SYSTEMS WITH APPLICATIONS 17 1.8994%
JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE AND TECHNOLOGY 16 1.7877%
ADVANCES IN KNOWLEDGE DISCOVERY AND DATA MINING, PROCEEDINGS 13 1.4525%
INFORMATION RETRIEVAL TECHNOLOGY, PROCEEDINGS 13 1.4525%
KNOWLEDGE AND INFORMATION SYSTEMS 13 1.4525%
JOURNAL OF MACHINE LEARNING RESEARCH 12 1.3408%
ADVANCED DATA MINING AND APPLICATIONS, PROCEEDINGS 11 1.2291%
COMPUTATIONAL LINGUISTICS AND INTELLIGENT TEXT PROCESSING 10 1.1173%
BRIEFINGS IN BIOINFORMATICS 9 1.0056%
NATURAL LANGUAGE PROCESSING AND INFORMATION SYSTEMS, PROCEEDINGS 9 1.0056%
ADVANCES IN ARTIFICIAL INTELLIGENCE, PROCEEDINGS 8 0.8939%
IEICE TRANSACTIONS ON INFORMATION AND SYSTEMS 7 0.7821%
SOFT COMPUTING 7 0.7821%
ADVANCES IN WEB-AGE INFORMATION MANAGEMENT, PROCEEDINGS 6 0.6704%
APPLIED SOFT COMPUTING 6 0.6704%
ARTIFICIAL INTELLIGENCE IN MEDICINE 6 0.6704%
DATABASE AND EXPERT SYSTEMS APPLICATIONS, PROCEEDINGS 6 0.6704%
DECISION SUPPORT SYSTEMS 6 0.6704%

DISCOVERY SCIENCE, PROCEEDINGS 6 0.6704%
FUZZY SYSTEMS AND KNOWLEDGE DISCOVERY, PT 2, PROCEEDINGS 6 0.6704%
INFORMATION RETRIEVAL 6 0.6704%
JOURNAL OF BIOMEDICAL INFORMATICS 6 0.6704%
JOURNAL OF UNIVERSAL COMPUTER SCIENCE 6 0.6704%
KNOWLEDGE-BASED INTELLIGENT INFORMATION AND ENGINEERING SYSTEMS, PT 2, PROCEEDINGS 6 0.6704%
NATURAL LANGUAGE PROCESSING - IJCNLP 2004 6 0.6704%
PATTERN RECOGNITION AND IMAGE ANALYSIS, PT 2, PROCEEDINGS 6 0.6704%
ACM TRANSACTIONS ON INFORMATION SYSTEMS 5 0.5587%
ADVANCES IN APPLIED ARTIFICIAL INTELLIGENCE, PROCEEDINGS 5 0.5587%
APPLIED ARTIFICIAL INTELLIGENCE 5 0.5587%
DATA WAREHOUSING AND KNOWLEDGE DISCOVERY, PROCEEDINGS 5 0.5587%
FRONTIERS OF WWW RESEARCH AND DEVELOPMENT - APWEB 2006, PROCEEDINGS 5 0.5587%
INFORMATION SCIENCES 5 0.5587%
JOURNAL OF THE AMERICAN MEDICAL INFORMATICS ASSOCIATION 5 0.5587%
KNOWLEDGE-BASED INTELLIGENT INFORMATION AND ENGINEERING SYSTEMS, PT 1, PROCEEDINGS 5 0.5587%
KNOWLEDGE-BASED INTELLIGENT INFORMATION AND ENGINEERING SYSTEMS, PT 3, PROCEEDINGS 5 0.5587%
TECHNOLOGICAL FORECASTING AND SOCIAL CHANGE 5 0.5587%
TEXT, SPEECH AND DIALOGUE, PROCEEDINGS 5 0.5587%
WEB TECHNOLOGIES RESEARCH AND DEVELOPMENT - APWEB 2005 5 0.5587%
ADVANCES IN NATURAL LANGUAGE PROCESSING 4 0.4469%
ADVANCES IN NATURAL LANGUAGE PROCESSING, PROCEEDINGS 4 0.4469%
AI 2004: ADVANCES IN ARTIFICIAL INTELLIGENCE, PROCEEDINGS 4 0.4469%

AI 2005: ADVANCES IN ARTIFICIAL INTELLIGENCE 4 0.4469%
COMPUTATIONAL AND INFORMATION SCIENCE, PROCEEDINGS 4 0.4469%
COMPUTATIONAL INTELLIGENCE AND SECURITY, PT 1, PROCEEDINGS 4 0.4469%
CONTENT COMPUTING, PROCEEDINGS 4 0.4469%
DATA & KNOWLEDGE ENGINEERING 4 0.4469%
DATA INTEGRATION IN THE LIFE SCIENCES, PROCEEDINGS 4 0.4469%
FOUNDATIONS OF INTELLIGENT SYSTEMS, PROCEEDINGS 4 0.4469%
FUZZY SYSTEMS AND KNOWLEDGE DISCOVERY, PROCEEDINGS 4 0.4469%
IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE 4 0.4469%
INFORMATION RETRIEVAL TECHNOLOGY 4 0.4469%
INNOVATIONS IN APPLIED ARTIFICIAL INTELLIGENCE 4 0.4469%
INTELLIGENCE AND SECURITY INFORMATICS, PROCEEDINGS 4 0.4469%
INTELLIGENT CONTROL AND AUTOMATION 4 0.4469%
INTELLIGENT DATA ENGINEERING AND AUTOMATED LEARNING - IDEAL 2006, PROCEEDINGS 4 0.4469%
INTERNATIONAL JOURNAL OF MEDICAL INFORMATICS 4 0.4469%
INTERNATIONAL JOURNAL ON ARTIFICIAL INTELLIGENCE TOOLS 4 0.4469%
KNOWLEDGE-BASED SYSTEMS 4 0.4469%
MACHINE LEARNING: ECML 2006, PROCEEDINGS 4 0.4469%
NATURAL LANGUAGE PROCESSING - IJCNLP 2005, PROCEEDINGS 4 0.4469%
NEURAL INFORMATION PROCESSING 4 0.4469%
NEUROCOMPUTING 4 0.4469%
PATTERN RECOGNITION AND MACHINE INTELLIGENCE, PROCEEDINGS 4 0.4469%
PROGRESS IN PATTERN RECOGNITION, IMAGE ANALYSIS AND APPLICATIONS, PROCEEDINGS 4 0.4469%
ADVANCES IN DATA MINING 3 0.3352%

ADVANCES IN NEURAL NETWORKS - ISNN 2006, PT 1 3 0.3352%
ADVANCES IN WEB INTELLIGENCE, PROCEEDINGS 3 0.3352%
ADVANCES IN WEB MINING AND WEB USAGE ANALYSIS 3 0.3352%
COMPUTATIONAL SCIENCE AND ITS APPLICATIONS - ICCSA 2005, PT 2 3 0.3352%
DATA MINING: THEORY, METHODOLOGY, TECHNIQUES, AND APPLICATIONS 3 0.3352%
DIGITAL LIBRARIES: IMPLEMENTING STRATEGIES AND SHARING EXPERIENCES, PROCEEDINGS 3 0.3352%
DIGITAL LIBRARIES: INTERNATIONAL COLLABORATION AND CROSS-FERTILIZATION, PROCEEDINGS 3 0.3352%
DOCUMENT ANALYSIS SYSTEMS VII, PROCEEDINGS 3 0.3352%
DRUG DISCOVERY TODAY 3 0.3352%
FUNDAMENTA INFORMATICAE 3 0.3352%
GRID AND COOPERATIVE COMPUTING GCC 2004, PROCEEDINGS 3 0.3352%
INFORMATION SYSTEMS 3 0.3352%
INTERNATIONAL JOURNAL OF INFORMATION TECHNOLOGY & DECISION MAKING 3 0.3352%
INTERNATIONAL JOURNAL OF MOLECULAR MEDICINE 3 0.3352%
JOURNAL OF COMPUTER SCIENCE AND TECHNOLOGY3 0.3352%
JOURNAL OF DATABASE MANAGEMENT 3 0.3352%
JOURNAL OF INFORMATION SCIENCE 3 0.3352%
JOURNAL OF INTELLIGENT INFORMATION SYSTEMS 3 0.3352%
JOURNAL OF MANAGEMENT INFORMATION SYSTEMS 3 0.3352%
KNOWLEDGE DISCOVERY IN DATABASES: PKDD 2005 3 0.3352%
KNOWLEDGE DISCOVERY IN LIFE SCIENCE LITERATURE, PROCEEDINGS 3 0.3352%
KNOWLEDGE SCIENCE, ENGINEERING AND MANAGEMENT 3 0.3352%

KNOWLEDGE-BASED INTELLIGENT INFORMATION AND ENGINEERING SYSTEMS, PT 4, PROCEEDINGS 3 0.3352%
MACHINE LEARNING 3 0.3352%
NEURAL INFORMATION PROCESSING, PT 3, PROCEEDINGS 3 0.3352%
NUCLEIC ACIDS RESEARCH 3 0.3352%
PLOS COMPUTATIONAL BIOLOGY 3 0.3352%
PROGRESS IN ARTIFICIAL INTELLIGENCE, PROCEEDINGS 3 0.3352%
SEMANTIC WEB - ASWC 2006, PROCEEDINGS 3 0.3352%
WEB INFORMATION SYSTEMS - WISE 2004, PROCEEDINGS 3 0.3352%
WEB INFORMATION SYSTEMS ENGINEERING - WISE 2005 3 0.3352%
ACTIVE MINING 2 0.2235%
ADVANCED WEB AND NETWORK TECHNOLOGIES, AND APPLICATIONS, PROCEEDINGS 2 0.2235%
ADVANCES IN INFORMATION RETRIEVAL 2 0.2235%
ADVANCES IN INTELLIGENT COMPUTING, PT 1, PROCEEDINGS 2 0.2235%
ADVANCES IN INTELLIGENT DATA ANALYSIS VI, PROCEEDINGS 2 0.2235%
ADVANCES IN MACHINE LEARNING AND CYBERNETICS 2 0.2235%
ADVANCES IN NATURAL COMPUTATION, PT 1, PROCEEDINGS 2 0.2235%
ADVANCES IN NEURAL NETWORKS - ISNN 2006, PT 2, PROCEEDINGS 2 0.2235%
ANNUAL REVIEW OF INFORMATION SCIENCE AND TECHNOLOGY 2 0.2235%
APPLIED INTELLIGENCE 2 0.2235%
ARTIFICIAL INTELLIGENCE: METHODOLOGY, SYSTEMS, AND APPLICATIONS, PROCEEDINGS 2 0.2235%
ARTIFICIAL NEURAL NETWORKS: FORMAL MODELS AND THEIR APPLICATIONS - ICANN 2005, PT 2, PROCEEDINGS 2 0.2235%
ASIST 2003: PROCEEDINGS OF THE 66TH ASIST ANNUAL MEETING, VOL 40, 2003 2 0.2235%
COMPARATIVE AND FUNCTIONAL GENOMICS 2 0.2235%
COMPUTATIONAL BIOLOGY AND CHEMISTRY 2 0.2235%

COMPUTATIONAL INTELLIGENCE 2 0.2235%
COMPUTATIONAL SCIENCE AND ITS APPLICATIONS - ICCSA 2006, PT 2 2 0.2235%
CURRENT TOPICS IN ARTIFICIAL INTELLIGENCE 2 0.2235%
DATABASES IN NETWORKED INFORMATION SYSTEMS, PROCEEDINGS 2 0.2235%
DISTRIBUTED COMPUTING AND INTERNET TECHNOLOGY, PROCEEDINGS 2 0.2235%
ENGINEERING APPLICATIONS OF ARTIFICIAL INTELLIGENCE 2 0.2235%
GENOME BIOLOGY 2 0.2235%
IEEE INTELLIGENT SYSTEMS 2 0.2235%
IEEE TRANSACTIONS ON SYSTEMS MAN AND CYBERNETICS PART A-SYSTEMS AND HUMANS 2 0.2235%
IEEE TRANSACTIONS ON SYSTEMS MAN AND CYBERNETICS PART B-CYBERNETICS 2 0.2235%
INTERNATIONAL JOURNAL OF APPROXIMATE REASONING 2 0.2235%
INTERNATIONAL JOURNAL OF COMPUTER MATHEMATICS 2 0.2235%
INTERNATIONAL JOURNAL OF ONCOLOGY 2 0.2235%
INTERNATIONAL JOURNAL OF PATTERN RECOGNITION AND ARTIFICIAL INTELLIGENCE 2 0.2235%
INTERNATIONAL JOURNAL OF SYSTEMS SCIENCE 2 0.2235%
JOURNAL OF BIOMEDICAL SCIENCE 2 0.2235%
JOURNAL OF DOCUMENTATION 2 0.2235%
JOURNAL OF WEB SEMANTICS 2 0.2235%
KNOWLEDGE DISCOVERY IN DATABASES: PKDD 2006, PROCEEDINGS 2 0.2235%
KNOWLEDGE EXPLORATION IN LIFE SCIENCE INFORMATICS, PROCEEDINGS 2 0.2235%
MACHINE LEARNING AND DATA MINING IN PATTERN RECOGNITION, PROCEEDINGS 2 0.2235%
MACHINE LEARNING: ECML 2005, PROCEEDINGS 2 0.2235%
ON THE MOVE TO MEANINGFUL INTERNET SYSTEMS 2005: OTM 2005 WORKSHOPS, PROCEEDINGS 2 0.2235%
ONLINE INFORMATION REVIEW 2 0.2235%
PARALLEL AND DISTRIBUTED COMPUTING: APPLICATIONS AND TECHNOLOGIES, PROCEEDINGS 2 0.2235%

PATTERN ANALYSIS AND APPLICATIONS 2 0.2235%
PATTERN RECOGNITION 2 0.2235%
PLOS BIOLOGY 2 0.2235%
PROCEEDINGS OF THE NATIONAL ACADEMY OF SCIENCES OF THE UNITED STATES OF AMERICA 2 0.2235%
RESEARCH AND ADVANCED TECHNOLOGY FOR DIGITAL LIBRARIES 2 0.2235%
ROUGH SETS AND KNOWLEDGE TECHNOLOGY, PROCEEDINGS 2 0.2235%
ROUGH SETS, FUZZY SETS, DATA MINING, AND GRANULAR COMPUTING, PT 2, PROCEEDINGS 2 0.2235%
SOCIOLOGICAL THEORY AND METHODS 2 0.2235%
STRING PROCESSING AND INFORMATION RETRIEVAL, PROCEEDINGS 2 0.2235%
TECHNOLOGY ANALYSIS & STRATEGIC MANAGEMENT 2 0.2235%
TECHNOMETRICS 2 0.2235%
TRANSACTIONS ON COMPUTATIONAL SYSTEMS BIOLOGY V 2 0.2235%
WEB INFORMATION SYSTEMS - WISE 2006 WORKSHOPS, PROCEEDINGS 2 0.2235%
WORLD WIDE WEB-INTERNET AND WEB INFORMATION SYSTEMS 2 0.2235%
ABSTRACTION, REFORMULATION AND APPROXIMATION, PROCEEDINGS 1 0.1117%
ACCESSING MULTILINGUAL INFORMATION REPOSITORIES 1 0.1117%
ACM COMPUTING SURVEYS 1 0.1117%
ADVANCED ENGINEERING INFORMATICS 1 0.1117%
ADVANCES IN ARTIFICIAL INTELLIGENCE - IBERAMIA 2004 1 0.1117%
ADVANCES IN DATABASE TECHNOLOGY - EDBT 2006 1 0.1117%
ADVANCES IN INFORMATICS, PROCEEDINGS 1 0.1117%
ADVANCES IN INTELLIGENT COMPUTING, PT 2, PROCEEDINGS 1 0.1117%
ADVANCES IN NATURAL COMPUTATION, PT 2, PROCEEDINGS 1 0.1117%
ADVANCES IN NEURAL NETWORKS - ISNN 2005, PT 1, PROCEEDINGS 1 0.1117%

ADVANCES IN NEURAL NETWORKS - ISNN 2005, PT 2, PROCEEDINGS 1 0.1117%
ADVANCES IN NEURAL NETWORKS - ISNN 2006, PT 3, PROCEEDINGS 1 0.1117%
ADVANCES IN XML INFORMATION RETRIEVAL 1 0.1117%
ADVANCES IN XML INFORMATION RETRIEVAL AND EVALUATION 1 0.1117%
AI COMMUNICATIONS 1 0.1117%
AI*IA2005: ADVANCES IN ARTIFICIAL INTELLIGENCE, PROCEEDINGS 1 0.1117%
ANALYTICAL BIOCHEMISTRY 1 0.1117%
APPLIED SPECTROSCOPY 1 0.1117%
ARTIFICIAL INTELLIGENCE AND SIMULATION 1 0.1117%
ARTIFICIAL INTELLIGENCE IN MEDICINE, PROCEEDINGS 1 0.1117%
ARTIFICIAL NEURAL NETWORKS: BIOLOGICAL INSPIRATIONS - ICANN 2005, PT 1, PROCEEDINGS 1 0.1117%
AUDIO AND VIDEO BASED BIOMETRIC PERSON AUTHENTICATION, PROCEEDINGS 1 0.1117%
AUTOMATION IN CONSTRUCTION 1 0.1117%
AUTONOMOUS INTELLIGENT SYSTEMS: AGENTS AND DATA MINING, PROCEEDINGS 1 0.1117%
BIOCHEMICAL SOCIETY TRANSACTIONS 1 0.1117%
BIOLOGICAL AND MEDICAL DATA ANALYSIS, PROCEEDINGS 1 0.1117%
BIOSYSTEMS 1 0.1117%
BIOTECHNOLOGY ADVANCES 1 0.1117%
BMC GENOMICS1 0.1117%
BONE 1 0.1117%
BT TECHNOLOGY JOURNAL 1 0.1117%
CANCER BIOLOGY & THERAPY 1 0.1117%
CELLULAR AND MOLECULAR LIFE SCIENCES 1 0.1117%
CELLULAR IMMUNOLOGY 1 0.1117%
CHINESE JOURNAL OF ELECTRONICS 1 0.1117%
COMMUNICATIONS OF THE ACM 1 0.1117%
COMPARATIVE BIOCHEMISTRY AND PHYSIOLOGY D-GENOMICS & PROTEOMICS 1 0.1117%
COMPARATIVE EVALUATION OF MULTILINGUAL INFORMATION ACCESS SYSTEMS 1 0.1117%

COMPONENT-BASED SOFTWARE ENGINEERING, PROCEEDINGS 1 0.1117%
COMPUTATIONAL INTELLIGENCE, PT 2, PROCEEDINGS 1 0.1117%
COMPUTATIONAL SCIENCE - ICCS 2005, PT 1, PROCEEDINGS 1 0.1117%
COMPUTATIONAL SCIENCE - ICCS 2006, PT 3, PROCEEDINGS 1 0.1117%
COMPUTER AND INFORMATION SCIENCES - ISCIS 2005, PROCEEDINGS 1 0.1117%
COMPUTER COMMUNICATIONS 1 0.1117%
COMPUTER MUSIC MODELING AND RETRIEVAL 1 0.1117%
COMPUTER NETWORKS 1 0.1117%
COMPUTER SCIENCE - THEORY AND APPLICATIONS 1 0.1117%
COMPUTERS & EDUCATION 1 0.1117%
COMPUTERS & GRAPHICS-UK 1 0.1117%
COMPUTERS & OPERATIONS RESEARCH 1 0.1117%
COMPUTERS ENVIRONMENT AND URBAN SYSTEMS 1 0.1117%
COMPUTING AND INFORMATICS 1 0.1117%
CONCEPTUAL MODELING - ER 2004, PROCEEDINGS 1 0.1117%
COOPERATIVE DESIGN, VISUALIZATION, AND ENGINEERING, PROCEEDINGS 1 0.1117%
COOPERATIVE INFORMATION AGENTS X, PROCEEDINGS 1 0.1117%
CRITICAL REVIEWS IN BIOTECHNOLOGY 1 0.1117%
CRITICAL REVIEWS IN MICROBIOLOGY 1 0.1117%
CURRENT NANOSCIENCE 1 0.1117%
CURRENT OPINION IN CHEMICAL BIOLOGY 1 0.1117%
CURRENT TRENDS IN DATABASE TECHNOLOGY - EDBT 2004 WORKSHOPS, PROCEEDINGS 1 0.1117%
CYBERNETICS AND SYSTEMS 1 0.1117%
DATA MINING AND KNOWLEDGE DISCOVERY 1 0.1117%
DATA MINING AND KNOWLEDGE MANAGEMENT 1 0.1117%
DATABASE SYSTEMS FOR ADVANCED APPLICATIONS, PROCEEDINGS 1 0.1117%

DYNAMICS OF CONTINUOUS DISCRETE AND IMPULSIVE SYSTEMS-SERIES B-APPLICATIONS & ALGORITHMS 1 0.1117%
E-COMMERCE AND WEB TECHNOLOGIES, PROCEEDINGS 1 0.1117%
ELECTRICAL ENGINEERING IN JAPAN 1 0.1117%
ELECTRONIC COMMERCE RESEARCH AND APPLICATIONS 1 0.1117%
EMERGING TRENDS IN INFORMATION AND COMMUNICATION SECURITY, PROCEEDINGS 1 0.1117%
ENERGY 1 0.1117%
EURASIP JOURNAL ON APPLIED SIGNAL PROCESSING 1 0.1117%
EUROPEAN JOURNAL OF HUMAN GENETICS 1 0.1117%
EUROPEAN JOURNAL OF OPERATIONAL RESEARCH 1 0.1117%
EXPERT SYSTEMS 1 0.1117%
FASEB JOURNAL 1 0.1117%
FEBS JOURNAL 1 0.1117%
FEDERATION OVER THE WEB 1 0.1117%
FLEXIBLE QUERY ANSWERING SYSTEMS, PROCEEDINGS 1 0.1117%
FUTURE GENERATION COMPUTER SYSTEMS 1 0.1117%
FUZZY SYSTEMS AND KNOWLEDGE DISCOVERY, PT 1, PROCEEDINGS 1 0.1117%
GENETIC ENGINEERING NEWS 1 0.1117%
GENETIC PROGRAMMING, PROCEEDINGS 1 0.1117%
GENOME RESEARCH 1 0.1117%
GLIA 1 0.1117%
GROUP DECISION AND NEGOTIATION 1 0.1117%
HEPATOLOGY 1 0.1117%
HIGH PERFORMANCE COMPUTING AND COMMUNICATIONS, PROCEEDINGS 1 0.1117%
HUMAN MOLECULAR GENETICS 1 0.1117%
HUMAN MUTATION 1 0.1117%
HUMAN-COMPUTER INTERACTION - INTERACT 2005, PROCEEDINGS 1 0.1117%
IBM SYSTEMS JOURNAL 1 0.1117%
IEEE ENGINEERING IN MEDICINE AND BIOLOGY MAGAZINE 1 0.1117%

IEEE INTERNET COMPUTING 1 0.1117%
IEEE TRANSACTIONS ON AUDIO SPEECH AND LANGUAGE PROCESSING 1 0.1117%
IEEE TRANSACTIONS ON INFORMATION TECHNOLOGY IN BIOMEDICINE 1 0.1117%
IEEE TRANSACTIONS ON MULTIMEDIA 1 0.1117%
IEEE TRANSACTIONS ON NANOBIOSCIENCE 1 0.1117%
IEEE TRANSACTIONS ON NEURAL NETWORKS 1 0.1117%
IEEE TRANSACTIONS ON SIGNAL PROCESSING 1 0.1117%
IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING 1 0.1117%
IEEE TRANSACTIONS ON SYSTEMS MAN AND CYBERNETICS PART C-APPLICATIONS AND REVIEWS 1 0.1117%
IEEE TRANSACTIONS ON VISUALIZATION AND COMPUTER GRAPHICS 1 0.1117%
IEEE-ACM TRANSACTIONS ON COMPUTATIONAL BIOLOGY AND BIOINFORMATICS 1 0.1117%
IEEE-ACM TRANSACTIONS ON COMPUTATIONAL BIOLOGY AND BIOINFORMATIOCS 1 0.1117%
IETE JOURNAL OF RESEARCH 1 0.1117%
INFORMATICA 1 0.1117%
INFORMATION SYSTEMS FRONTIERS 1 0.1117%
INNOVATIVE INTERNET COMMUNITY SYSTEMS 1 0.1117%
INTELLIGENT DATA ANALYSIS 1 0.1117%
INTELLIGENT TECHNIQUES FOR WEB PERSONALIZATION 1 0.1117%
INTELLIGENT TUTORING SYSTEMS, PROCEEDINGS 1 0.1117%
INTERACTING WITH COMPUTERS 1 0.1117%
INTERNATIONAL JOURNAL OF BIOCHEMISTRY & CELL BIOLOGY 1 0.1117%
INTERNATIONAL JOURNAL OF HUMAN-COMPUTER STUDIES 1 0.1117%
INTERNATIONAL JOURNAL OF INTELLIGENT SYSTEMS 1 0.1117%
INTERNATIONAL JOURNAL OF PRODUCTION RESEARCH 1 0.1117%
INTERNATIONAL JOURNAL OF TECHNOLOGY MANAGEMENT 1 0.1117%

INTERNATIONAL JOURNAL OF WEB SERVICES RESEARCH 1 0.1117%
INTERNET RESEARCH 1 0.1117%
IRANIAN JOURNAL OF SCIENCE AND TECHNOLOGY TRANSACTION B-ENGINEERING 1 0.1117%
JOURNAL OF AIRCRAFT 1 0.1117%
JOURNAL OF ARTIFICIAL INTELLIGENCE RESEARCH 1 0.1117%
JOURNAL OF BIOLOGICAL CHEMISTRY 1 0.1117%
JOURNAL OF COMPUTATIONAL BIOLOGY 1 0.1117%
JOURNAL OF INTELLIGENT & FUZZY SYSTEMS 1 0.1117%
JOURNAL OF NANOPARTICLE RESEARCH 1 0.1117%
JOURNAL OF NETWORK AND COMPUTER APPLICATIONS 1 0.1117%
JOURNAL OF NEW MUSIC RESEARCH 1 0.1117%
JOURNAL OF NUCLEAR SCIENCE AND TECHNOLOGY 1 0.1117%
JOURNAL OF PROTEOME RESEARCH 1 0.1117%
JOURNAL OF PUBLIC HEALTH 1 0.1117%
JOURNAL OF SOCIOLINGUISTICS 1 0.1117%
JOURNAL OF SYSTEMS ARCHITECTURE 1 0.1117%
JOURNAL OF THE CHINESE INSTITUTE OF ENGINEERS 1 0.1117%
JOURNAL ON DATA SEMANTICS II 1 0.1117%
KNOWLEDGE DISCOVERY FROM XML DOCUMENTS, PROCEEDINGS 1 0.1117%
KNOWLEDGE ENGINEERING REVIEW 1 0.1117%
KUWAIT JOURNAL OF SCIENCE & ENGINEERING 1 0.1117%
LANGUAGE RESOURCES AND EVALUATION 1 0.1117%
LIBRARY AND INFORMATION SCIENCE 1 0.1117%
LIPIDS IN HEALTH AND DISEASE 1 0.1117%
LOCAL PATTERN DETECTION 1 0.1117%
LOGIC JOURNAL OF THE IGPL 1 0.1117%
LOGIC PROGRAMMING, PROCEEDINGS 1 0.1117%
MASSIVELY MULTI-AGENT SYSTEMS I 1 0.1117%
MECHANISMS OF AGEING AND DEVELOPMENT 1 0.1117%
METHODS OF INFORMATION IN MEDICINE 1 0.1117%
MICAI 2005: ADVANCES IN ARTIFICIAL INTELLIGENCE1 0.1117%

MODELING DECISIONS FOR ARTIFICIAL INTELLIGENCE 1 0.1117%
MOLECULAR CARCINOGENESIS 1 0.1117%
NATURE REVIEWS GENETICS 1 0.1117%
NEC TECHNICAL JOURNAL 1 0.1117%
NETWORKING AND MOBILE COMPUTING, PROCEEDINGS 1 0.1117%
NEURAL COMPUTATION 1 0.1117%
NEURAL INFORMATION PROCESSING, PT 2, PROCEEDINGS 1 0.1117%
NEURAL NETWORKS 1 0.1117%
NEXT GENERATION INFORMATION TECHNOLOGIES AND SYSTEMS, PROCEEDINGS 1 0.1117%
OMEGA-INTERNATIONAL JOURNAL OF MANAGEMENT SCIENCE 1 0.1117%
OMICS-A JOURNAL OF INTEGRATIVE BIOLOGY 1 0.1117%
ON THE CONVERGENCE OF BIO-INFORMATION-, ENVIRONMENTAL-, ENERGY-, SPACE- AND NANO-TECHNOLOGIES, PTS 1 AND 2 1 0.1117%
ONLINE 1 0.1117%
PHYSIOLOGICAL GENOMICS 1 0.1117%
PLANT PHYSIOLOGY 1 0.1117%
PRACTICAL ASPECTS OF KNOWLEDGE MANAGEMENT, PROCEEDINGS 1 0.1117%
PRICAI 2006: TRENDS IN ARTIFICIAL INTELLIGENCE, PROCEEDINGS 1 0.1117%
PRINCIPLES AND PRACTICE OF SEMANTIC WEB REASONING 1 0.1117%
PRODUCT FOCUSED SOFTWARE PROCESS IMPROVEMENT, PROCEEDINGS 1 0.1117%
PROTEOMICS 1 0.1117%
QSAR & COMBINATORIAL SCIENCE 1 0.1117%
REASONING WEB 1 0.1117%
REVISTA SIGNOS 1 0.1117%
SADHANA-ACADEMY PROCEEDINGS IN ENGINEERING SCIENCES 1 0.1117%
SCIENCE AND ENGINEERING ETHICS 1 0.1117%
SCIENTOMETRICS 1 0.1117%
SEMANTIC WEB - ISWC 2005, PROCEEDINGS1 0.1117%

SEMANTIC WEB: RESEARCH AND APPLICATIONS, PROCEEDINGS 1 0.1117%
SOFTWARE-PRACTICE & EXPERIENCE 1 0.1117%
SPATIAL INFORMATION THEORY, PROCEEDINGS 1 0.1117%
SPEECH COMMUNICATION 1 0.1117%
SUPPORTIVE CARE IN CANCER 1 0.1117%
TECHNOLOGY IN CANCER RESEARCH & TREATMENT 1 0.1117%
THEORETICAL COMPUTER SCIENCE 1 0.1117%
TOURISM MANAGEMENT 1 0.1117%
TRENDS IN BIOTECHNOLOGY 1 0.1117%
USER MODELING 2005, PROCEEDINGS 1 0.1117%
USER MODELING AND USER-ADAPTED INTERACTION 1 0.1117%
VLDB JOURNAL 1 0.1117%
WEB AND COMMUNICATION TECHNOLOGIES AND INTERNET - RELATED SOCIAL ISSUES - HSI 2005 1 0.1117%
WEB INFORMATION SYSTEMS ENGINEERING - WISE 2005 WORKSHOPS, PROCEEDINGS 1 0.1117%

IMPORTANT PHRASES RELATED TO TEXT MINING
(2005 AND PRIOR YEARS)

BIOINFORMATICS
TEXT CATEGORIZATION
TEXT MINING
TEXT CLASSIFICATION
INFORMATION RETRIEVAL
DOCUMENT CLUSTERING
FEATURE SELECTION
INFORMATION EXTRACTION
DOCUMENT CLASSIFICATION
DOCUMENT COLLECTIONS
SEARCH ENGINES
CLUSTERING ALGORITHMS
SELF-ORGANIZING MAP
HIERARCHICAL CLUSTERING
TEXT CLASSIFIER
TEXT CLUSTERING
TEXT RETRIEVAL
DOCUMENT CLUSTERS
DOCUMENT RETRIEVAL

TERM CLUSTERING
DOCUMENT CATEGORIZATION
TERM SELECTION
SELF-ORGANIZING MAPS
DOCUMENT CLUSTER
CLASSIFICATION ALGORITHMS
FEATURE EXTRACTION
RETRIEVED DOCUMENTS
LITERATURE MINING
RETRIEVE DOCUMENTS
WEB MINING
KEYWORD CLUSTERS
LATENT SEMANTIC INDEXING
NATURAL LANGUAGE PROCESSING
VECTOR SPACE MODEL
SINGULAR VALUE DECOMPOSITION
SUPPORT VECTOR MACHINE
PRECISION AND RECALL
INFORMATION RETRIEVAL SYSTEMS
AUTOMATIC TEXT CATEGORIZATION
DOCUMENT CLUSTERING ALGORITHM
TEXT DATA MINING
TEXTUAL DATA MINING
WORD SENSE DISAMBIGUATION
AUTOMATED TEXT CATEGORIZATION
AUTOMATIC TEXT CLASSIFICATION
TEXT CATEGORIZATION SYSTEM
TEXT CLASSIFICATION SYSTEM
AGGLOMERATIVE HIERARCHICAL CLUSTERING
MACHINE LEARNING ALGORITHMS
HIERARCHICAL CLUSTERING ALGORITHM
MULTILINGUAL TEXT CATEGORIZATION
WEB DOCUMENT CLUSTERING
AUTOMATIC DOCUMENT CLASSIFICATION
DOCUMENT CLUSTERING ALGORITHMS
EXTRACTING MULTI-WORD PHRASE
MULTI-WORD PHRASE FREQUENCIES
AUTOMATED TEXT CLASSIFICATION
CLUSTERING OF DOCUMENTS
HIERARCHICAL TEXT CLASSIFICATION
INFORMATION EXTRACTION SYSTEMS
INTERNET SEARCH ENGINES
LATENT SEMANTIC STRUCTURE
WEB SEARCH ENGINE
TEXT CLASSIFIERS

PROLIFIC AUTHORS CONTRIBUTING TO TEXT MINING

KOSTOFF--RN
KATOH--M
SWANSON--DR
BERRY--MW
SMALHEISER--NR
LEE--CH
MONTANES--E
COMBARRO--EF
DIAZ--I
RANILLA--J
BI--YX
CHEN--HC
GORDON--MD
MA--FY
SRINIVASAN--P
YANG--HC
ZHANG--J
FELDMAN--R
FUKETA--M
LIU--F
MERKL--D
TOOTHMAN--DR
WEISS--SM
YANG--YM
BUNKE--H
DENG--ZH
FERNANDEZ--J
HSU--CC
HUMENIK--JA
KARYPIS--G
KLOPOTEK--MA
KOSTER--CHA
LAM--W
LAST--M
LI--ML
MONS--B
MORITA--K
RUIZ-SHULCLOPER--J
SCHENKER--A
SUN--MS
TANG--SW
THEERAMUNKONG--T
WEEBER--M
YIN--HJ
ZHANG--M
AOE--J
ATLAM--ES

AUMANN--Y
BELL--D
CHEN--WL
CHIANG--JH
CHIEN--LF
DE MOOR--B
DEL RIO--JA
EBERHART--HJ
GELBUKH--A
KADOYA--Y
KAMEL--MS
KANDEL--A
KIM--HJ
KORS--JA
LEE--GG
LERTNATTEE--V
LIU--H
LOPEZ-LOPEZ--A
LUO--X
MONTES-Y-GOMEZ--M
PARK--H
SCHNEIDER--KM
SEBASTIANI--F
WANG--DL
WEI--CP
WERMTER--S
YAO--TS
ZHANG--T
ZHU--JB

MOST CITED FIRST AUTHORS

KATOH M
SALTON G
YANG Y
KOSTOFF RN
JOACHIMS T
LEWIS DD
SWANSON DR
DUMAIS ST
KOHONEN T
DEERWESTER S
SEBASTIANI F
MCCALLUM A
HEARST MA
BERRY MW
VANRIJSBERGEN CJ
APTE C

COHEN WW
VAPNIK VN
FELDMAN R
SMALHEISER NR
NIGAM K
AGRAWAL R
CHEN HC
QUINLAN JR
MITCHELL TM
PORTER MF
ZAMIR O
JAIN AK
MERKL D
NARIN F
MLADENIC D
SCHAPIRE RE
CRAVEN M
CHAKRABARTI S
DHILLON IS
LAM W
YANG YM
BAEZAYATES R
RAUBER A
SCHUTZE H
VOORHEES EM
LAGUS K
CUTTING DR
WILLETT P
HONKELA T
FREUND Y
KASKI S
BLASCHKE C
FRIEDMAN C
LANDAUER TK
SAHAMI M
WEISS SM
DUDA RO
ROCCHIO JJ
KOLLER D
KARYPIS G
BRILL E
BUCKLEY C
HARMAN D
RILOFF E
HOFMANN T
WITTEN IH
GOLUB GH
GORDON MD

LIN X
WEEBER M
FUHR N
NG HT
WIENER E
SCHOLKOPF B
ALLAN J
BLUM A
BOLEY D
MANNING CD
CRISTIANINI N
HAN J
KIRIKOSHI H
RINDFLESCH TC
ROBERTSON SE
CORTES C
CROFT WB
FRAKES WB
JENSSEN TK
LANG K
PUSTEJOVSKY J
SAITOH T
SINGHAL A
ZHANG T
DEMPSTER AP
FELLBAUM C
SMALL H
BAKER LD
BEZDEK JC
FRITZKE B
FURNAS GW
HULL D
SRINIVASAN P
CALLON M
COVER TM
FUKUDA K
RASMUSSEN E
STEINBACH M
DOMINGOS P
GARFIELD E
SLONIM N
DAGAN I
GUHA S
HATZIVASSILOGLO.V
MILLER GA
STREHL A
TANABE L
WERMTER S

JONES KS
SHATKAY H
AGGARWAL CC
KOLDA TG
MCCALLUM AK
RUIZ ME
ZHAO Y

JOURNALS CONTAINING TEXT MINING ARTICLES

INFORMATION PROCESSING & MANAGEMENT
JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE AND TECHNOLOGY
COMPUTATIONAL LINGUISTICS AND INTELLIGENT TEXT PROCESSING
IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING
BIOINFORMATICS
INFORMATION RETRIEVAL
ADVANCES IN INFORMATION RETRIEVAL
MACHINE LEARNING
KNOWLEDGE-BASED INTELLIGENT INFORMATION AND ENGINEERING SYSTEMS, PT 1, PROCEEDINGS
JOURNAL OF INTELLIGENT INFORMATION SYSTEMS
SCIENTOMETRICS
ACM TRANSACTIONS ON INFORMATION SYSTEMS
PRINCIPLES OF DATA MINING AND KNOWLEDGE DISCOVERY
APPLIED INTELLIGENCE
IEICE TRANSACTIONS ON INFORMATION AND SYSTEMS
INFORMATION RETRIEVAL TECHNOLOGY, PROCEEDINGS
PATTERN RECOGNITION LETTERS
IBM SYSTEMS JOURNAL
ADVANCES IN WEB-AGE INFORMATION MANAGEMENT, PROCEEDINGS
PATTERN RECOGNITION
NEUROCOMPUTING
NATURAL LANGUAGE PROCESSING AND INFORMATION SYSTEMS, PROCEEDINGS
NATURAL LANGUAGE PROCESSING - IJCNLP 2004
INFORMATION SCIENCES
BMC BIOINFORMATICS
INTERNATIONAL JOURNAL OF MOLECULAR MEDICINE
ADVANCES IN KNOWLEDGE DISCOVERY AND DATA MINING, PROCEEDINGS
DECISION SUPPORT SYSTEMS
ARTIFICIAL INTELLIGENCE
IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE
INTERNATIONAL JOURNAL OF INTELLIGENT SYSTEMS
EXPERT SYSTEMS WITH APPLICATIONS

KNOWLEDGE-BASED INTELLIGENT INFORMATION AND ENGINEERING SYSTEMS, PT 3, PROCEEDINGS
JOURNAL OF THE AMERICAN MEDICAL INFORMATICS ASSOCIATION
INTELLIGENCE AND SECURITY INFORMATICS, PROCEEDINGS
ADVANCED DATA MINING AND APPLICATIONS, PROCEEDINGS
APPLIED ARTIFICIAL INTELLIGENCE
PATTERN RECOGNITION AND IMAGE ANALYSIS, PT 2, PROCEEDINGS
KNOWLEDGE DISCOVERY IN DATABASES: PKDD 2004, PROCEEDINGS
FOUNDATIONS OF INTELLIGENT SYSTEMS, PROCEEDINGS
PERSPECTIVES IN BIOLOGY AND MEDICINE
ADVANCES IN WEB-AGE INFORMATION MANAGEMENT: PROCEEDINGS
INTERNATIONAL JOURNAL OF MEDICAL INFORMATICS
INTERNATIONAL JOURNAL OF ONCOLOGY
TECHNOLOGICAL FORECASTING AND SOCIAL CHANGE
JOURNAL OF BIOMEDICAL INFORMATICS
DATA WAREHOUSING AND KNOWLEDGE DISCOVERY, PROCEEDINGS
DISCOVERY SCIENCE, PROCEEDINGS
KNOWLEDGE DISCOVERY IN DATABASES: PKDD 2005
DATABASE AND EXPERT SYSTEMS APPLICATIONS, PROCEEDINGS
NEURAL INFORMATION PROCESSING
DATA MINING AND KNOWLEDGE DISCOVERY
JOURNAL OF CHEMICAL INFORMATION AND COMPUTER SCIENCES
NEURAL NETWORKS
FOUNDATIONS OF INTELLIGENT SYSTEMS
NATURAL LANGUAGE PROCESSING AND INFORMATION SYSTEMS
BRIEFINGS IN BIOINFORMATICS
CONTENT COMPUTING, PROCEEDINGS
WEB INFORMATION SYSTEMS - WISE 2004, PROCEEDINGS
JOURNAL OF MACHINE LEARNING RESEARCH
JOURNAL OF MANAGEMENT INFORMATION SYSTEMS
FUZZY SYSTEMS AND KNOWLEDGE DISCOVERY, PT 2, PROCEEDINGS
IEEE INTELLIGENT SYSTEMS
ARTIFICIAL INTELLIGENCE IN MEDICINE
COMPUTERS AND THE HUMANITIES
GRID AND COOPERATIVE COMPUTING GCC 2004, PROCEEDINGS
JOURNAL OF UNIVERSAL COMPUTER SCIENCE
KNOWLEDGE AND INFORMATION SYSTEMS
JOURNAL OF INFORMATION SCIENCE
INNOVATIONS IN APPLIED ARTIFICIAL INTELLIGENCE
COMPUTATIONAL LINGUISTICS AND INTELLIGENT TEXT PROCESSING, PROCEEDINGS
INTERNATIONAL JOURNAL OF APPROXIMATE REASONING
INTELLIGENT DATA ENGINEERING AND AUTOMATED LEARNING - IDEAL 2002
INTELLIGENT DATA ENGINEERING AND AUTOMATED LEARNING
ADVANCES IN INFORMATION RETRIEVAL, PROCEEDINGS

ADVANCED WEB TECHNOLOGIES AND APPLICATIONS
COMPUTATIONAL AND INFORMATION SCIENCE, PROCEEDINGS
SIAM JOURNAL ON MATRIX ANALYSIS AND APPLICATIONS
ADVANCES IN NATURAL LANGUAGE PROCESSING

PROLIFIC INSTITUTIONS CONTRIBUTING TO TEXT MINING

OFF NAVAL RES
IBM CORP
TSING HUA UNIV
UNIV CHICAGO
SHANGHAI JIAO TONG UNIV
M&M MED BIOINFORMAT
NATL CANC CTR
NATL UNIV SINGAPORE
CARNEGIE MELLON UNIV
UNIV ILLINOIS
UNIV TOKYO
UNIV TENNESSEE
UNIV MINNESOTA
CNR
CHINESE ACAD SCI
UNIV CALIF BERKELEY
NORTHEASTERN UNIV
UNIV OVIEDO
NANYANG TECHNOL UNIV
QUEENS UNIV BELFAST
NATL INST INFORMAT
UNIV IOWA
PEKING UNIV
UNIV MARYLAND
UNIV ARIZONA
BEN GURION UNIV NEGEV
UNIV TEXAS
UNIV TOKUSHIMA
UNIV ULSTER
SEOUL NATL UNIV
PENN STATE UNIV
NATL KAOHSIUNG UNIV APPL SCI
NATL TAIWAN UNIV
UNIV MICHIGAN
ZHEJIANG UNIV
RUTGERS STATE UNIV
UNIV QUEENSLAND
UNIV TSUKUBA
BAR ILAN UNIV
STANFORD UNIV
UNIV N CAROLINA

CHINESE UNIV HONG KONG
UNIV NIJMEGEN
CHANG JUNG UNIV
INDIAN INST TECHNOL
FUDAN UNIV
UNIV WATERLOO
UNIV WAIKATO
NOESIS INC
UNIV KARLSRUHE
UNIV POLITECN VALENCIA
UNIV PITTSBURGH
UNIV LONDON ROYAL HOLLOWAY & BEDFORD NEW COLL
KATHOLIEKE UNIV LEUVEN
HELSINKI UNIV TECHNOL
UNIV S FLORIDA
UNIV GRANADA
KOREA ADV INST SCI & TECHNOL
NAGOYA INST TECHNOL
COLUMBIA UNIV
POLISH ACAD SCI
POHANG UNIV SCI & TECHNOL
UNIV SO CALIF
UNIV BERN
UNIV CALIF LOS ANGELES
UNIV COLL LONDON
ATHENS UNIV ECON & BUSINESS
NATL CHENG KUNG UNIV
UNIV MASSACHUSETTS
INST CYBERNET MATH & PHYS
USN
UNIV ORIENTE
UNIV PASSAU
MONASH UNIV
KYOTO UNIV
UNIV MANCHESTER
ACAD SINICA
XIAN JIAOTONG UNIV
THAMMASAT UNIV
TOKYO INST TECHNOL
UNIV LEIPZIG
UNIV WASHINGTON
UNIV EDINBURGH
GEORGE MASON UNIV
NATL TAIWAN UNIV SCI & TECHNOL
NATL SUN YAT SEN UNIV
UNIV SUNDERLAND

PROLIFIC COUNTRIES CONTRIBUTING TO TEXT MINING

USA
PEOPLES R CHINA
JAPAN
GERMANY
SPAIN
SOUTH KOREA
ENGLAND
TAIWAN
CANADA
AUSTRALIA
ITALY
FRANCE
SINGAPORE
NETHERLANDS
ISRAEL
INDIA
GREECE
SWITZERLAND
FINLAND
SCOTLAND
NORTH IRELAND
MEXICO
BELGIUM
NEW ZEALAND
THAILAND
POLAND
BRAZIL
SWEDEN
AUSTRIA
HUNGARY
CUBA
PORTUGAL
RUSSIA
SLOVENIA
TURKEY

TEXT MINING BIBLIOGRAPHY

[Anon]. 2005. Feature selection improves text classification accuracy. IEEE INTELLIGENT SYSTEMS 20 (6): 75-75.
Abdullah, MT; Ahmad, F; Mahmod, R; Sembok, TMT. 2003. Application of latent semantic indexing on Malay-English cross language information retrieval. DIGITAL LIBRARIES: TECHNOLOGY AND MANAGEMENT

OF INDIGENOUS KNOWLEDGE FOR GLOBAL ACCESS 2911: 663-665. LECTURE NOTES IN COMPUTER SCIENCE
Abulaish, M; Dey, L. 2005. An ontology-based pattern mining system for extracting information from biological texts. ROUGH SETS, FUZZY SETS, DATA MINING, AND GRANULAR COMPUTING, PT 2, PROCEEDINGS 3642: 420-429. LECTURE NOTES IN ARTIFICIAL INTELLIGENCE
Adak, S. 2002. e2eXpress: End-to-end bioinformatics and knowledge management system for microarrays. JOURNAL OF BIOLOGICAL SYSTEMS 10 (4): 285-302.
Adami, G; Avesani, P; Sona, D. 2005. Clustering documents into a web directory for bootstrapping a supervised classification. DATA & KNOWLEDGE ENGINEERING 54 (3): 301-325.
Aggarwal, CC; Gates, SC; Yu, PS. 2004. On using partial supervision for text categorization. IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING 16 (2): 245-255.
Aggarwal, CC; Yu, PS. 1999. On text mining techniques for personalization. NEW DIRECTIONS IN ROUGH SETS, DATA MINING, AND GRANULAR-SOFT COMPUTING 1711: 12-18. LECTURE NOTES IN ARTIFICIAL INTELLIGENCE
Agrawal, R; Ho, H; Jacquenet, F; Jacquenet, M. 2005. Mining information extraction rules from datasheets without linguistic parsing. INNOVATIONS IN APPLIED ARTIFICIAL INTELLIGENCE 3533: 510-520. LECTURE NOTES IN ARTIFICIAL INTELLIGENCE
Aguilar-Ruiz, JS; Rodriguez-Baena, DS; Cohen, PR; Riquelme, JC. 2003. Clustering main concepts from e-mails. CURRENT TOPICS IN ARTIFICIAL INTELLIGENCE 3040: 231-240. LECTURE NOTES IN COMPUTER SCIENCE
AGUIRRE, LA. 1994. TERM CLUSTERING AND THE ORDER SELECTION OF LINEAR CONTINUOUS SYSTEMS. JOURNAL OF THE FRANKLIN INSTITUTE-ENGINEERING AND APPLIED MATHEMATICS 331B (4): 403-415.
AGUIRRE, LA; BILLINGS, SA. 1995. IMPROVED STRUCTURE SELECTION FOR NONLINEAR MODELS BASED ON TERM CLUSTERING. INTERNATIONAL JOURNAL OF CONTROL 62 (3): 569-587.
Ahmad, K; Vrusias, BL; Ledford, A. 2001. Choosing feature sets for training and testing self-organising maps: A case study. NEURAL COMPUTING & APPLICATIONS 10 (1): 56-66.

Ahonen, H; Heinonen, O; Klemettinen, M; Verkamo, AI. 1997. Mining in the phrasal frontier. PRINCIPLES OF DATA MINING AND KNOWLEDGE DISCOVERY 1263: 343-350. LECTURE NOTES IN ARTIFICIAL INTELLIGENCE
Aihara, K; Takasu, A. 2001. Category based customization approach for information retrieval. USER MODELING 2001, PROCEEDINGS 2109: 207-209. LECTURE NOTES IN ARTIFICIAL INTELLIGENCE
Aizawa, A. 2003. An information-theoretic perspective of tf-idf measures. INFORMATION PROCESSING & MANAGEMENT 39 (1): 45-65.
Albert, S; Gaudan, S; Knigge, H; Raetsch, A; Delgado, A; Huhse, B; Kirsch, H; Albers, M; Rebholz-Schuhmann, D; Koegl, M. 2003. Computer-assisted generation of a protein-interaction database for nuclear receptors. MOLECULAR ENDOCRINOLOGY 17 (8): 1555-1567.
Alexandrov, VN; Dimov, IT; Karaivanova, A; Tan, CJK. 2003. Parallel Monte Carlo algorithms for information retrieval. MATHEMATICS AND COMPUTERS IN SIMULATION 62 (3-6): 289-295.
Alfonseca, E; Perez, D; Rodriguez, P. 2004. WELKIN: Automatic generation of adaptive hypermedia sites with NLP techniques. WEB ENGINEERING, PROCEEDINGS 3140: 617-618. LECTURE NOTES IN COMPUTER SCIENCE
Allan, J; Leuski, A; Swan, R; Byrd, D. 2001. Evaluating combinations of ranked lists and visualizations of inter-document similarity. INFORMATION PROCESSING & MANAGEMENT 37 (3): 435-458.
Almpanidis, G; Kotropoulos, C. 2005. Combining text and link analysis for focused crawling. PATTERN RECOGNITION AND DATA MINING, PT 1, PROCEEDINGS 3686: 278-287. LECTURE NOTES IN COMPUTER SCIENCE
Ampazis, N; Perantonis, SJ. 2004. LSISOM - A latent semantic indexing approach to Self-Organizing Maps of document collections. NEURAL PROCESSING LETTERS 19 (2): 157-173.
Antal, P; Fannes, G; Timmerman, D; Moreau, Y; De Moor, B. 2004. Using literature and data to learn Bayesian networks as clinical models of ovarian tumors. ARTIFICIAL INTELLIGENCE IN MEDICINE 30 (3): 257-281.
Aphinyanaphongs, Y; Tsamardinos, I; Statnikov, A; Hardin, D; Aliferis, CF. 2005. Text categorization models for high-quality article retrieval in internal medicine. JOURNAL OF THE AMERICAN MEDICAL INFORMATICS ASSOCIATION 12 (2): 207-216.
APTE, C; DAMERAU, F; WEISS, SM. 1994. AUTOMATED LEARNING OF DECISION RULES FOR TEXT CATEGORIZATION. ACM TRANSACTIONS ON INFORMATION SYSTEMS 12 (3): 233-251.

Apte, CV; Hong, SJ; Natarajan, R; Pednault, EPD; Tipu, FA; Weiss, SM. 2003. Data-intensive analytics for predicting modeling. IBM JOURNAL OF RESEARCH AND DEVELOPMENT 47 (1): 17-23.
Arevian, G; Wermter, S; Panchev, C. 2003. Symbolic state transducers and recurrent neural preference machines for text mining. INTERNATIONAL JOURNAL OF APPROXIMATE REASONING 32 (2-3): 237-258.
Arimura, H. 2002. Efficient text mining with optimized pattern discovery. COMBINATORIAL PATTERN MATCHING 2373: 17-19. LECTURE NOTES IN COMPUTER SCIENCE
Atkinson-Abutridy, J. 2004. Semantically-driven explanatory text mining: Beyond keywords. ADVANCES IN ARTIFICIAL INTELLIGENCE - IBERAMIA 2004 3315: 275-285. LECTURE NOTES IN ARTIFICIAL INTELLIGENCE
Atkinson-Abutridy, J; Mellish, C; Aitken, S. 2003. A semantically guided and domain-independent evolutionary model for knowledge discovery from texts. IEEE TRANSACTIONS ON EVOLUTIONARY COMPUTATION 7 (6): 546-560.
Atkinson-Abutridy, J; Mellish, C; Aitken, S. 2004. Combining information extraction with genetic algorithms for text mining. IEEE INTELLIGENT SYSTEMS 19 (3): 22-30.
Aumann, Y; Feldman, R; Ben Yehuda, Y; Landau, D; Liphstat, O; Schler, Y. 1999. Circle graphs: New visualization tools for text-mining. PRINCIPLES OF DATA MINING AND KNOWLEDGE DISCOVERY 1704: 277-282. LECTURE NOTES IN ARTIFICIAL INTELLIGENCE
Azcarraga, AP; Yap, TN; Tan, J; Chua, TS. 2004. Evaluating keyword selection methods for WEBSOM text archives. IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING 16 (3): 380-383.
Bajdik, CD; Kuo, B; Rusaw, S; Jones, S; Brooks-Wilson, A. 2005. CGMIM: Automated text-mining of Online Mendelian Inheritance in Man (OMIM) to identify genetically-associated cancers and candidate genes. BMC BIOINFORMATICS 6: art. no.-78.
Bajic, VB; Veronika, M; Veladandi, PS; Meka, A; Heng, MW; Rajaraman, K; Pan, H; Swarup, S. 2005. Dragon plant biology explorer. A text-mining tool for integrating associations between genetic and biochemical entities with genome annotation and biochemical terms lists. PLANT PHYSIOLOGY 138 (4): 1914-1925.
Banerjee, A; Ghosh, J. 2004. Frequency-sensitive competitive learning for scalable balanced clustering on high-dimensional hyperspheres. IEEE TRANSACTIONS ON NEURAL NETWORKS 15 (3): 702-719.

Bansal, N; Blum, A; Chawla, S. 2004. Correlation clustering. MACHINE LEARNING 56 (1-3): 89-113.
Bao, JP; Shen, JY; Liu, HY; Liu, XD. 2006. A fast document copy detection model. SOFT COMPUTING 10 (1): 41-46.
Bao, JP; Shen, JY; Liu, XD; Liu, HY; Zhang, XD. 2004. Finding plagiarism based on common semantic sequence model. ADVANCES IN WEB-AGE INFORMATION MANAGEMENT: PROCEEDINGS 3129: 640-645. LECTURE NOTES IN COMPUTER SCIENCE
Bao, YG; Asai, D; Du, XY; Ishii, N. 2003. Partition for the rough set-based text classification. ADVANCES IN WEB-AGE INFORMATION MANAGEMENT, PROCEEDINGS 2762: 181-188. LECTURE NOTES IN COMPUTER SCIENCE
Bao, YG; Asai, D; Du, XY; Yamada, K; Ishii, N. 2003. An effective rough set-based method for text classification. INTELLIGENT DATA ENGINEERING AND AUTOMATED LEARNING 2690: 545-552. LECTURE NOTES IN COMPUTER SCIENCE
Bao, YG; Ishii, N. 2002. Combining multiple k-nearest neighbor classifiers for text classification by reducts. DISCOVERY SCIENCE, PROCEEDINGS 2534: 340-347. LECTURE NOTES IN COMPUTER SCIENCE
BARTELL, BT; COTTRELL, GW; BELEW, RK. 1995. REPRESENTING DOCUMENTS USING AN EXPLICIT MODEL OF THEIR SIMILARITIES. JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE 46 (4): 254-271.
Basili, R; Cammisa, M; Moschitti, A. 2005. A semantic kernel to exploit linguistic knowledge. AI*IA2005: ADVANCES IN ARTIFICIAL INTELLIGENCE, PROCEEDINGS 3673: 290-302. LECTURE NOTES IN ARTIFICIAL INTELLIGENCE
Basili, R; DellaRocca, M; Pazienza, MT. 1997. Contextual word sense tuning and disambiguation. APPLIED ARTIFICIAL INTELLIGENCE 11 (3): 235-262.
Bayer, T; Kressel, U; Mogg-Schneider, H; Renz, I. 1998. Categorizing paper documents - A generic system for domain and language independent text categorization. COMPUTER VISION AND IMAGE UNDERSTANDING 70 (3): 299-306.
Begeja, L; Drucker, H; Gibbon, D; Haffner, P; Liu, Z; Renger, B; Shahraray, B. 2005. Semantic data mining of short utterances. IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING 13 (5): 672-680, Part 1.

Behme, W; Mucksch, H. 1999. Selection and classification of external information for the integration in a data warehouse. WIRTSCHAFTSINFORMATIK 41 (5): 443-+.
Bel, N; Koster, CHA; Villegas, M. 2003. Cross-Lingual Text Categorization. RESEARCH AND ADVANCED TECHNOLOGY FOR DIGITAL LIBRARIES 2769: 126-139. LECTURE NOTES IN COMPUTER SCIENCE
Belkin, M; Niyogi, P. 2004. Semi-supervised learning on Riemannian manifolds. MACHINE LEARNING 56 (1-3): 209-239.
Bell, DA; Guan, JW; Bi, YX. 2005. On combining classifier mass functions for text categorization. IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING 17 (10): 1307-1319.
Bender, HJ. 2002. Natural intelligence in a machine translation system. MACHINE TRANSLATION: FROM RESEARCH TO REAL USERS 2499: 224-228. LECTURE NOTES IN ARTIFICIAL INTELLIGENCE
Bengel, J; Gauch, S; Mittur, E; Vijayaraghavan, R. 2004. ChatTrack: Chat room topic detection using classification. INTELLIGENCE AND SECURITY INFORMATICS, PROCEEDINGS 3073: 266-277. LECTURE NOTES IN COMPUTER SCIENCE
Benkhalifa, M; Mouradi, A; Bouyakhf, H. 2001. Integrating external knowledge to supplement training data in semi-supervised learning for text categorization. INFORMATION RETRIEVAL 4 (2): 91-113.
Benkhalifa, M; Mouradi, A; Bouyakhf, H. 2001. Integrating WordNet knowledge to supplement training data in semi-supervised agglomerative hierarchical clustering for text categorization. INTERNATIONAL JOURNAL OF INTELLIGENT SYSTEMS 16 (8): 929-947.
Bennett, PN; Dumais, ST; Horvitz, E. 2005. The combination of text classifiers using reliability indicators. INFORMATION RETRIEVAL 8 (1): 67-100.
Berardi, M; Lapi, M; Leo, P; Loglisci, C. 2005. Mining generalized association rules on biomedical literature. INNOVATIONS IN APPLIED ARTIFICIAL INTELLIGENCE 3533: 500-509. LECTURE NOTES IN ARTIFICIAL INTELLIGENCE
Berger, H; Merkl, D. 2004. A comparison of text-categorization methods applied to N-gram frequency statistics. AI 2004: ADVANCES IN ARTIFICIAL INTELLIGENCE, PROCEEDINGS 3339: 998-1003. LECTURE NOTES IN ARTIFICIAL INTELLIGENCE
Bergman, CM; Carlson, JW; Celniker, SE. 2005. Drosophila DNase I footprint database: a systematic genome annotation of transcription factor

binding sites in the fruitfly, Drosophila melanogaster. BIOINFORMATICS 21 (8): 1747-1749.
Berry, MW; Dumais, ST; OBrien, GW. 1995. Using linear algebra for intelligent information retrieval. SIAM REVIEW 37 (4): 573-595.
Berry, MW; Fierro, RD. 1996. Low-rank orthogonal decompositions for information retrieval applications. NUMERICAL LINEAR ALGEBRA WITH APPLICATIONS 3 (4): 301-327.
Berry, MW; Pulatova, SA; Stewart, GW. 2005. Algorithm 844: Computing sparse reduced-rank approximations to sparse matrices. ACM TRANSACTIONS ON MATHEMATICAL SOFTWARE 31 (2): 252-269.
Berry, MW; Young, PG. 1995. Using latent semantic indexing for multilanguage information retrieval. COMPUTERS AND THE HUMANITIES 29 (6): 413-429.
Berryman, MJ; Allison, A; Abbott, D. 2003. Statistical techniques for text classification based on word recurrence intervals. FLUCTUATION AND NOISE LETTERS 3 (1): L1-L10.
Betister, G; Furbach, U; Gross-Hardt, M; Thomas, B. 2003. Automatic classification for the identification of relationships in a meta-data repository. DISCOVERY SCIENCE, PROCEEDINGS 2843: 283-290. LECTURE NOTES IN ARTIFICIAL INTELLIGENCE
Bi, YX; Anderson, T; McClean, S. 2004. Combining rules for text categorization using Dempster's rule of combination. INTELLIGENT DATA ENGINEERING AND AUTOMATED LEARNING IDEAL 2004, PROCEEDINGS 3177: 457-463. LECTURE NOTES IN COMPUTER SCIENCE
Bi, YX; Anderson, T; McClean, S. 2004. Multiple sets of rules for text categorization. ADVANCES IN INFORMATION SYSTEMS, PROCEEDINGS 3261: 263-272. LECTURE NOTES IN COMPUTER SCIENCE
Bi, YX; Bell, D; Guan, JW. 2004. Combining evidence from classifiers in text categorization. KNOWLEDGE-BASED INTELLIGENT INFORMATION AND ENGINEERING SYSTEMS, PT 3, PROCEEDINGS 3215: 521-528. LECTURE NOTES IN ARTIFICIAL INTELLIGENCE
Bi, YX; Bell, D; Wang, H; Gu, GG; Greer, K. 2004. Combining multiple classifiers using Dempster's rule of combination for text categorization. MODELING DECISIONS FOR ARTIFICIAL INTELLIGENCE, PROCEEDINGS 3131: 127-138. LECTURE NOTES IN ARTIFICIAL INTELLIGENCE

Bi, YX; Bell, D; Wang, H; Guo, GD; Dubitzky, W. 2004. Classification decision combination for text categorization: An experimental study. DATABASE AND EXPERT SYSTEMS APPLICATIONS, PROCEEDINGS 3180: 222-231. LECTURE NOTES IN COMPUTER SCIENCE
Biagioli, C; Francesconi, E; Spinosa, P; Taddei, M. 2004. XML documents within a legal domain: Standards and tools for the Italian legislative environment. DOCUMENT ANALYSIS SYSTEMS VI, PROCEEDINGS 3163: 413-424. LECTURE NOTES IN COMPUTER SCIENCE
Biemann, C; Quasthoff, U; Bohm, K; Wolff, C. 2003. Automatic discovery and aggregation of compound names for the use in knowledge representations. JOURNAL OF UNIVERSAL COMPUTER SCIENCE 9 (6): 530-541.
Bigi, B. 2003. Using Kullback-Leibler distance for text categorization. ADVANCES IN INFORMATION RETRIEVAL 2633: 305-319. LECTURE NOTES IN COMPUTER SCIENCE
Blaschke, C; Hirschman, L; Yeh, A; Valencia, A. 2003. Critical assessment of information extraction systems in biology. COMPARATIVE AND FUNCTIONAL GENOMICS 4 (6): 674-677.
Blei, DM; Ng, AY; Jordan, MI. 2003. Latent Dirichlet allocation. JOURNAL OF MACHINE LEARNING RESEARCH 3 (4-5): 993-1022.
Blom, K; Ruhe, A. 2004. A Krylov subspace method for information retrieval. SIAM JOURNAL ON MATRIX ANALYSIS AND APPLICATIONS 26 (2): 566-582.
Bogg, P. 2003. Pattern based approaches to pre-processing structured text: A newsfeed example. COMPUTATIONAL SCIENCE - ICCS 2003, PT IV, PROCEEDINGS 2660: 859-867. LECTURE NOTES IN COMPUTER SCIENCE
Bohm, K; Heyer, G; Quasthoff, U; Wolff, C. 2002. Topic map generation using text mining. JOURNAL OF UNIVERSAL COMPUTER SCIENCE 8 (6): 623-633.
Bolioli, A; Mercatali, P; Romano, F. 2004. Formal models for a legislative grammar - Explicit text amendment. KNOWLEDGE MANAGEMENT IN ELECTRONIC GOVERNMENT, PROCEEDINGS 3025: 194-211. LECTURE NOTES IN COMPUTER SCIENCE
Bordogna, G; Pasi, G. 1996. A user-adaptive neural network supporting a rule-based relevance feedback. FUZZY SETS AND SYSTEMS 82 (2): 201-211.

Borzemski, L; Lopatka, P. 2005. Complementing search engines with text mining. INNOVATIONS IN APPLIED ARTIFICIAL INTELLIGENCE 3533: 743-745. LECTURE NOTES IN ARTIFICIAL INTELLIGENCE
Bote, VPG; Anegon, FD; Solana, VH. 2002. Document organization using Kohonen's algorithm. INFORMATION PROCESSING & MANAGEMENT 38 (1): 79-89.
Boutell, MR; Luo, JB; Shen, XP; Brown, CM. 2004. Learning multi-label scene classification. PATTERN RECOGNITION 37 (9): 1757-1771.
Bradford, RB. 2005. Efficient discovery of new information in large text databases. INTELLIGENCE AND SECURITY INFORMATICS, PROCEEDINGS 3495: 374-380. LECTURE NOTES IN COMPUTER SCIENCE
Bratko, A; Filipic, B. 2006. Exploiting structural information for semi-structured document categorization. INFORMATION PROCESSING & MANAGEMENT 42 (3): 679-694.
Braun, T; Schubert, A; Kostoff, RN. 2002. A chemistry field in search of applications statistical analysis of US fullerene patents. JOURNAL OF CHEMICAL INFORMATION AND COMPUTER SCIENCES 42 (5): 1011-1015.
Bremer, EG; Natarajan, J; Zhang, YH; DeSesa, C; Hack, CJ; Dubitzky, W. 2004. Text mining of full text articles and creation of a knowledge base for analysis of microarray data. KNOWLEDGE EXPLORATION IN LIFE SCIENCE INFORMATICS, PROCEEDINGS 3303: 84-95. LECTURE NOTES IN ARTIFICIAL INTELLIGENCE
Buckeridge, AM; Sutcliffe, RFE. 2002. Using latent semantic indexing as a measure of conceptual association for noun compound disambiguation. ARTIFICIAL INTELLIGENCE AND COGNITIVE SCIENCE, PROCEEDINGS 2464: 12-19. LECTURE NOTES IN ARTIFICIAL INTELLIGENCE
Bunescu, R; Ge, RF; Kate, RJ; Marcotte, EM; Mooney, RJ; Ramani, AK; Wong, YW. 2005. Comparative experiments on learning information extractors for proteins and their interactions. ARTIFICIAL INTELLIGENCE IN MEDICINE 33 (2): 139-155.
BURGIN, R. 1995. THE RETRIEVAL EFFECTIVENESS OF 5 CLUSTERING ALGORITHMS AS A FUNCTION OF INDEXING EXHAUSTIVITY. JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE 46 (8): 562-572.
Burstein, J; Marcu, D. 2003. A machine learning approach for identification of thesis and conclusion statements in student essays. COMPUTERS AND THE HUMANITIES 37 (4): 455-467.

Cai, D; He, XF; Han, JW. 2005. Document clustering using locality preserving indexing. IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING 17 (12): 1624-1637.
Cairns, P. 2004. Informalising formal mathematics: Searching the Mizar library with latent semantics. MATHEMATICAL KNOWLEDGE MANAGEMENT, PROCEEDINGS 3119: 58-72. LECTURE NOTES IN COMPUTER SCIENCE
Caldas, CH; Soibelman, L. 2003. Automating hierarchical document classification for construction management information systems. AUTOMATION IN CONSTRUCTION 12 (4): 395-406.
Caldas, CH; Soibelman, L; Han, JW. 2002. Automated classification of construction project documents. JOURNAL OF COMPUTING IN CIVIL ENGINEERING 16 (4): 234-243.
Camon, E; Magrane, M; Barrell, D; Lee, V; Dimmer, E; Maslen, J; Binns, D; Harte, N; Lopez, R; Apweiler, R. 2004. The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology. NUCLEIC ACIDS RESEARCH 32: D262-D266, Sp. Iss. SI.
CAN, F; OZKARAHAN, EA. 1990. CONCEPTS AND EFFECTIVENESS OF THE COVER-COEFFICIENT-BASED CLUSTERING METHODOLOGY FOR TEXT DATABASES. ACM TRANSACTIONS ON DATABASE SYSTEMS 15 (4): 483-517.
Cardoso-Cachopo, A; Oliveira, AL. 2003. An empirical comparison of text categorization methods. STRING PROCESSING AND INFORMATION RETRIEVAL, PROCEEDINGS 2857: 183-196. LECTURE NOTES IN COMPUTER SCIENCE
Cascini, G; Fantechi, A; Spinicci, E. 2004. Natural language processing of patents and technical documentation. DOCUMENT ANALYSIS SYSTEMS VI, PROCEEDINGS 3163: 508-520. LECTURE NOTES IN COMPUTER SCIENCE
Cascini, G; Rissone, P. 2004. Plastics design: integrating TRIZ creativity and semantic knowledge portals. JOURNAL OF ENGINEERING DESIGN 15 (4): 405-424.
Casillas, A; de Lena, MTG; Martinez, R. 2003. Document clustering into an unknown number of clusters using a genetic algorithm. TEXT, SPEECH AND DIALOGUE, PROCEEDINGS 2807: 43-49. LECTURE NOTES IN ARTIFICIAL INTELLIGENCE
Casillas, A; de Lena, MTG; Martinez, R. 2004. Sampling and feature selection in a genetic algorithm for document clustering. COMPUTATIONAL LINGUISTICS AND INTELLIGENT TEXT

PROCESSING 2945: 601-612. LECTURE NOTES IN COMPUTER SCIENCE
Cesa-Bianchi, N; Conconi, A; Gentile, C. 2003. Learning probabilistic linear-threshold classifiers via selective sampling. LEARNING THEORY AND KERNEL MACHINES 2777: 373-387. LECTURE NOTES IN ARTIFICIAL INTELLIGENCE
Chakrabarti, S; Roy, S; Soundalgekar, MV. 2003. Fast and accurate text classification via multiple linear discriminant projections. VLDB JOURNAL 12 (2): 170-185.
Chali, Y; Noureddine, S. 2005. Document clustering with grouping and chaining algorithms. NATURAL LANGUAGE PROCESSING - IJCNLP 2005, PROCEEDINGS 3651: 280-291. LECTURE NOTES IN ARTIFICIAL INTELLIGENCE
Chang, CH; Hsu, CC. 1998. Integrating query expansion and conceptual relevance feedback for personalized Web information retrieval. COMPUTER NETWORKS AND ISDN SYSTEMS 30 (1-7): 621-623.
Chang, CH; Hsu, CC. 1999. Enabling concept-based relevance feedback for information retrieval on the WWW. IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING 11 (4): 595-609.
Chang, HC; Hsu, CC. 2005. Using topic keyword clusters for automatic document clustering. IEICE TRANSACTIONS ON INFORMATION AND SYSTEMS E88D (8): 1852-1860.
Chang, HC; Hsu, CC; Chan, CK. 2004. Automatic document clustering based on keyword clusters using partitions of weighted diagraphs. COMPUTER SYSTEMS SCIENCE AND ENGINEERING 19 (1): 27-37.
Chapman, WW; Bridewell, W; Hanbury, P; Cooper, GF; Buchanan, BG. 2001. A simple algorithm for identifying negated findings and diseases in discharge summaries. JOURNAL OF BIOMEDICAL INFORMATICS 34 (5): 301-310.
Chapman, WW; Christensen, LM; Wagner, MM; Haug, PJ; Ivanov, O; Dowling, JN; Olszewski, RT. 2005. Classifying free-text triage chief complaints into syndromic categories with natural language processing. ARTIFICIAL INTELLIGENCE IN MEDICINE 33 (1): 31-40.
Chau, R; Yeh, CH. 2003. Fuzzy methods for knowledge discovery from multilingual text. KNOWLEDGE-BASED INTELLIGENT INFORMATION AND ENGINEERING SYSTEMS, PT 1, PROCEEDINGS 2773: 835-842. LECTURE NOTES IN ARTIFICIAL INTELLIGENCE
Chau, RN; Yeh, CS; Smith, KA. 2005. A neural network model for hierarchical multilingual text categorization. ADVANCES IN NEURAL

NETWORKS - ISNN 2005, PT 2, PROCEEDINGS 3497: 238-245. LECTURE NOTES IN COMPUTER SCIENCE
Chau, RW; Yeh, CH. 2004. A multilingual text mining approach to web cross-lingual text retrieval. KNOWLEDGE-BASED SYSTEMS 17 (5-6): 219-227.
Chen, B. 2006. Exploring the use of latent topical information for statistical Chinese spoken document retrieval. PATTERN RECOGNITION LETTERS 27 (1): 9-18.
Chen, CM. 1999. Visualising semantic spaces and author co-citation networks in digital libraries. INFORMATION PROCESSING & MANAGEMENT 35 (3): 401-420.
Chen, CM; Lee, HM; Hwang, CW. 2005. A hierarchical neural network document classifier with linguistic feature selection. APPLIED INTELLIGENCE 23 (3): 277-294.
Chen, CS; Hiscott, RN. 1999. Statistical analysis of facies clustering in submarine-fan turbidite successions. JOURNAL OF SEDIMENTARY RESEARCH 69 (2): 505-517, Part B.
Chen, DY; Li, X. 2004. PLD: A distillation algorithm for misclassified documents. ADVANCES IN WEB-AGE INFORMATION MANAGEMENT: PROCEEDINGS 3129: 499-508. LECTURE NOTES IN COMPUTER SCIENCE
Chen, DY; Li, X; Dong, ZY; Chen, X. 2005. Effectiveness of document representation for classification. DATA WAREHOUSING AND KNOWLEDGE DISCOVERY, PROCEEDINGS 3589: 368-377. LECTURE NOTES IN COMPUTER SCIENCE
Chen, H; Sharp, BM. 2004. Content-rich biological network constructed by mining PubMed abstracts. BMC BIOINFORMATICS 5: art. no.-147.
Chen, HC; Chau, M; Zeng, D. 2002. CI Spider: a tool for competitive intelligence on the Web. DECISION SUPPORT SYSTEMS 34 (1): 1-17.
Chen, HC; Fan, HY; Chau, M; Zeng, D. 2003. Testing a Cancer Meta Spider. INTERNATIONAL JOURNAL OF HUMAN-COMPUTER STUDIES 59 (5): 755-776.
Chen, HH; Kuo, JJ; Su, TC. 2003. Clustering and visualization in a multi-lingual multi-document summarization system. ADVANCES IN INFORMATION RETRIEVAL 2633: 266-280. LECTURE NOTES IN COMPUTER SCIENCE
Chen, J; Yin, J; Zhang, J; Huang, J. 2005. Associative classification in text categorization. ADVANCES IN INTELLIGENT COMPUTING, PT 1, PROCEEDINGS 3644: 1035-1044. LECTURE NOTES IN COMPUTER SCIENCE

Chen, L; Huang, J; Gong, ZH. 2005. An anti-noise text categorization method based on support vector machines. ADVANCES IN WEB INTELLIGENCE, PROCEEDINGS 3528: 272-278. LECTURE NOTES IN COMPUTER SCIENCE
Chen, L; Tokuda, N; Nagai, A. 2003. A new differential LSI space-based probabilistic document classifier. INFORMATION PROCESSING LETTERS 88 (5): 203-212.
Chen, LH; Chue, WL. 2005. Using Web structure and summarisation techniques for Web content mining. INFORMATION PROCESSING & MANAGEMENT 41 (5): 1225-1242.
Chen, WL; Chang, XZ; Wang, HH; Zhu, JB; Yao, TS. 2005. Automatic word clustering for text categorization using global information. INFORMATION RETRIEVAL TECHNOLOGY 3411: 1-11. LECTURE NOTES IN COMPUTER SCIENCE
Chen, WL; Zhu, JB; Wu, HL; Yao, TS. 2004. Automatic learning features using bootstrapping for text categorization. COMPUTATIONAL LINGUISTICS AND INTELLIGENT TEXT PROCESSING 2945: 571-579. LECTURE NOTES IN COMPUTER SCIENCE
Chen, XY; Chen, Y; Li, RL; Hu, YF. 2005. An improvement of text association classification using rules weights. ADVANCED DATA MINING AND APPLICATIONS, PROCEEDINGS 3584: 355-363. LECTURE NOTES IN ARTIFICIAL INTELLIGENCE
Cheung, Z; Phan, KL; Mahidadia, A; Hoffmann, A. 2004. Feature extraction for learning to classify questions. AI 2004: ADVANCES IN ARTIFICIAL INTELLIGENCE, PROCEEDINGS 3339: 1069-1075. LECTURE NOTES IN ARTIFICIAL INTELLIGENCE
Chiang, JH; Chen, YC. 2004. An intelligent news recommender agent for filtering and categorizing large volumes of text corpus. INTERNATIONAL JOURNAL OF INTELLIGENT SYSTEMS 19 (3): 201-216.
Chiang, JH; Yu, HC. 2003. MeKE: discovering the functions of gene products from biomedical literature via sentence alignment. BIOINFORMATICS 19 (11): 1417-1422.
Chiang, JH; Yu, HC. 2005. Literature extraction of protein functions using sentence pattern mining. IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING 17 (8): 1088-1098.
Chiang, JH; Yu, HC; Hsu, HJ. 2004. GIS: a biomedical text-mining system for gene information discovery. BIOINFORMATICS 20 (1): 120-121.
Chik, FCY; Luk, RWP; Chung, KFL. 2005. Text categorization based on subtopic clusters. NATURAL LANGUAGE PROCESSING AND

INFORMATION SYSTEMS, PROCEEDINGS 3513: 203-214. LECTURE NOTES IN COMPUTER SCIENCE
Cho, SB; Lee, JH. 2003. Learning neural network ensemble for practical text classification. INTELLIGENT DATA ENGINEERING AND AUTOMATED LEARNING 2690: 1032-1036. LECTURE NOTES IN COMPUTER SCIENCE
Choi, B; Guo, Q. 2003. Applying semantic links for classifying Web pages. DEVELOPMENTS IN APPLIED ARTIFICIAL INTELLIGENCE 2718: 148-153. LECTURE NOTES IN ARTIFICIAL INTELLIGENCE
Chouchoulas, A; Shen, Q. 1999. A rough set-based approach to text classification. NEW DIRECTIONS IN ROUGH SETS, DATA MINING, AND GRANULAR-SOFT COMPUTING 1711: 118-127. LECTURE NOTES IN ARTIFICIAL INTELLIGENCE
Chouchoulas, A; Shen, Q. 2001. Rough set-aided keyword reduction for text categorization. APPLIED ARTIFICIAL INTELLIGENCE 15 (9): 843-873.
Chowdhury, N; Saha, D. 2005. Unsupervised text classification using Kohonen's Self Organizing Network. COMPUTATIONAL LINGUISTICS AND INTELLIGENT TEXT PROCESSING 3406: 715-718. LECTURE NOTES IN COMPUTER SCIENCE
Chu, M; Del Buono, N; Lopez, L; Politi, T. 2005. On the low-rank approximation of data on the unit sphere. SIAM JOURNAL ON MATRIX ANALYSIS AND APPLICATIONS 27 (1): 46-60.
Chuang, SL; Chien, LF. 2005. Taxonomy generation for text segments: A practical Web-based approach. ACM TRANSACTIONS ON INFORMATION SYSTEMS 23 (4): 363-396.
Chuang, W; Parker, DS. 2001. Pyramidal digest: An efficient model for abstracting text databases. DATABASE AND EXPERT SYSTEMS APPLICATIONS 2113: 360-369. LECTURE NOTES IN COMPUTER SCIENCE
Chuang, WT; Tiyyagura, A; Yang, J; Giuffrida, G. 2000. A fast algorithm for hierarchical text classification. DATA WAREHOUSING AND KNOWLEDGE DISCOVERY, PROCEEDINGS 1874: 409-418. LECTURE NOTES IN COMPUTER SCIENCE
Chung, S; McLeod, D. 2003. Dynamic topic mining from news stream data. ON THE MOVE TO MEANINGFUL INTERNET SYSTEMS 2003: COOPIS, DOA, AND ODBASE 2888: 653-670. LECTURE NOTES IN COMPUTER SCIENCE
Chung, S; McLeod, D. 2005. Dynamic pattern mining: An incremental data clustering approach. JOURNAL ON DATA SEMANTICS II 3360: 85-112. LECTURE NOTES IN COMPUTER SCIENCE

Chung, YM; Lee, JY. 2001. A corpus-based approach to comparative evaluation of statistical term association measures. JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE AND TECHNOLOGY 52 (4): 283-296.
CHUTE, CG; YANG, Y. 1995. AN OVERVIEW OF STATISTICAL-METHODS FOR THE CLASSIFICATION AND RETRIEVAL OF PATIENT EVENTS. METHODS OF INFORMATION IN MEDICINE 34 (1-2): 104-110.
Civera, J; Cubel, E; Juan, A; Vidal, E. 2005. Different approaches to bilingual text classification based on grammatical inference techniques. PATTERN RECOGNITION AND IMAGE ANALYSIS, PT 2, PROCEEDINGS 3523: 630-637. LECTURE NOTES IN COMPUTER SCIENCE
Cody, WF; Kreulen, JT; Krishna, V; Spangler, WS. 2002. The integration of business intelligence and knowledge management. IBM SYSTEMS JOURNAL 41 (4): 697-713.
Cohen, AM; Hersh, WR; Dubay, C; Spackman, K. 2005. Using co-occurrence network structure to extract synonymous gene and protein names from MEDLINE abstracts. BMC BIOINFORMATICS 6: art. no.-103.
Cohen, G; Hilario, M; Pellegrini, C. 2004. One-class support vector machines with a conformal kernel. A case study in handling class imbalance. STRUCTURAL, SYNTACTIC, AND STATISTICAL PATTERN RECOGNITION, PROCEEDINGS 3138: 850-858. LECTURE NOTES IN COMPUTER SCIENCE
Cohen, WW. 2000. WHIRL: A word-based information representation language. ARTIFICIAL INTELLIGENCE 118 (1-2): 163-196.
Cohen, WW; Singer, Y. 1999. Context-sensitive learning methods for text categorization. ACM TRANSACTIONS ON INFORMATION SYSTEMS 17 (2): 141-173.
Collier, N; Takeuchi, K. 2004. Comparison of character-level and part of speech features for name recognition in biomedical texts. JOURNAL OF BIOMEDICAL INFORMATICS 37 (6): 423-435.
Collier, N; Takeuchi, K; Kawazoe, A; Mullen, T; Wattarujeekrit, T. 2003. A framework for integrating deep and shallow semantic structures in text mining. KNOWLEDGE-BASED INTELLIGENT INFORMATION AND ENGINEERING SYSTEMS, PT 1, PROCEEDINGS 2773: 824-834. LECTURE NOTES IN ARTIFICIAL INTELLIGENCE
Collins-Thompson, K; Callan, J. 2005. Predicting reading difficulty with statistical language models. JOURNAL OF THE AMERICAN SOCIETY

FOR INFORMATION SCIENCE AND TECHNOLOGY 56 (13): 1448-1462.
Combarro, EF; Montanes, E; Diaz, I; Ranilla, J; Mones, R. 2005. Introducing a family of linear measures for feature selection in text categorization. IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING 17 (9): 1223-1232.
Combarro, EF; Montanes, E; Ranilla, J; Fernandez, J. 2003. A comparison of the performance of SVM and ARNI on Text Categorization with new filtering measures on an unbalanced collection. ARTIFICIAL NEURAL NETS PROBLEM SOLVING METHODS, PT II 2687: 742-749. LECTURE NOTES IN COMPUTER SCIENCE
Cong, G; Lee, WS; Wu, HR; Liu, B. 2004. Semi-supervised text classification using partitioned EM. DATABASE SYSTEMS FOR ADVANCED APPLICATIONS 2973: 482-493. LECTURE NOTES IN COMPUTER SCIENCE
Cordon, O; Herrera-Viedma, E; Lopez-Pujalte, C; Luque, M; Zarco, C. 2003. A review on the application of evolutionary computation to information retrieval. INTERNATIONAL JOURNAL OF APPROXIMATE REASONING 34 (2-3): 241-264.
Corney, DPA; Buxton, BF; Langdon, WB; Jones, DT. 2004. BioRAT: extracting biological information from full-length papers. BIOINFORMATICS 20 (17): 3206-3213.
Corral, A. 2004. Long-term clustering, scaling, and universality in the temporal occurrence of earthquakes. PHYSICAL REVIEW LETTERS 92 (10): art. no.-108501.
Correa, RF; Ludermir, TB. 2004. Dimensionality reduction by semantic mapping in text categorization. NEURAL INFORMATION PROCESSING 3316: 1032-1037. LECTURE NOTES IN COMPUTER SCIENCE
Correa, RF; Ludermir, TB. 2004. Web documents categorization using neural networks. NEURAL INFORMATION PROCESSING 3316: 758-762. LECTURE NOTES IN COMPUTER SCIENCE
Coutinho, DP; Figueiredo, MAT. 2005. Information theoretic text classification using the Ziv-Merhav method. PATTERN RECOGNITION AND IMAGE ANALYSIS, PT 2, PROCEEDINGS 3523: 355-362. LECTURE NOTES IN COMPUTER SCIENCE
Crasto, CJ; Marenco, LN; Migliore, M; Mao, BQ; Nadkarni, PM; Miller, P; Shepherd, GM. 2003. Text mining neuroscience journal articles to populate neuroscience databases. NEUROINFORMATICS 1 (3): 215-237.

Craven, M; DiPasquo, D; Freitag, D; McCallum, A; Mitchell, T; Nigam, K; Slattery, S. 2000. Learning to construct knowledge bases from the World Wide Web. ARTIFICIAL INTELLIGENCE 118 (1-2): 69-113.
Craven, M; Slattery, S. 2001. Relational learning with statistical predicate invention: Better models for hypertext. MACHINE LEARNING 43 (1-2): 97-119.
Cristianini, N; Scholkopf, B. 2002. Support vector machines and kernel methods - The new generation of learning machines. AI MAGAZINE 23 (3): 31-41.
Cristianini, N; Shawe-Taylor, J; Lodhi, H. 2002. Latent semantic kernels. JOURNAL OF INTELLIGENT INFORMATION SYSTEMS 18 (2-3): 127-152.
Cumbo, C; Iiritano, S; Rullo, P. 2004. OLEX - A reasoning-based text classifier. LOGICS IN ARTIFICIAL INTELLIGENCE, PROCEEDINGS 3229: 722-725. LECTURE NOTES IN COMPUTER SCIENCE
Cumbo, C; Iiritano, S; Rullo, P. 2004. Reasoning-based knowledge extraction for text classification. DISCOVERY SCIENCE, PROCEEDINGS 3245: 380-387. LECTURE NOTES IN COMPUTER SCIENCE
Damerau, FJ; Zhang, T; Weiss, SM; Indurkhya, N. 2004. Text categorization for a comprehensive time-dependent benchmark. INFORMATION PROCESSING & MANAGEMENT 40 (2): 209-221.
Danger, R; Berlanga, R; Ruiz-Shulcloper, J. 2004. CRISOL: An approach for automatically populating Semantic Web from unstructured text collections. DATABASE AND EXPERT SYSTEMS APPLICATIONS, PROCEEDINGS 3180: 243-252. LECTURE NOTES IN COMPUTER SCIENCE
Danger, R; Ruiz-Shulcloper, J; Berlanga, R. 2003. Text mining using the hierarchical syntactical structure of documents. CURRENT TOPICS IN ARTIFICIAL INTELLIGENCE 3040: 556-565. LECTURE NOTES IN COMPUTER SCIENCE
Dasigi, V; Mann, RC; Protopopescu, VA. 2001. Information fusion for text classification - an experimental comparison. PATTERN RECOGNITION 34 (12): 2413-2425.
Davies, NJ; Weeks, R; Revett, MC. 1996. Information agents for the World Wide Web. BT TECHNOLOGY JOURNAL 14 (4): 105-114.
de Bruijn, B; Martin, J. 2002. Getting to the (c)ore of knowledge: mining biomedical literature. INTERNATIONAL JOURNAL OF MEDICAL INFORMATICS 67 (1-3): 7-18.
De Bruijn, B; Martin, J; Wolting, C; Donaldson, I. 2001. Extracting sentences to justify categorization. ASIST 2001: PROCEEDINGS OF THE

64TH ASIST ANNUAL MEETING, VOL 38, 2001 38: 450-457. PROCEEDINGS OF THE ASIST ANNUAL MEETING
de Campos, LM; Fernandez-Luna, JM; Huete, JF. 2004. Clustering terms in the Bayesian network retrieval model: a new approach with two term-layers. APPLIED SOFT COMPUTING 4 (2): 149-158.
De Comite, F; Gilleron, R; Tommasi, M. 2003. Learning multi-label alternating decision trees from texts and data. MACHINE LEARNING AND DATA MINING IN PATTERN RECOGNITION, PROCEEDINGS 2734: 35-49. LECTURE NOTES IN ARTIFICIAL INTELLIGENCE
De Moor, B; Marchal, K; Mathys, J; Moreau, Y. 2003. Bioinformatics: Organisms from Venus, technology from Jupiter, algorithms from Mars. EUROPEAN JOURNAL OF CONTROL 9 (2-3): 237-278.
de Oliveira, JPM; Loh, S; Wives, LK; Scarinci, RG; Musa, D; Silva, L; Zambenedetti, C. 2004. Applying text mining on electronic messages for competitive intelligence. E-COMMERCE AND WEB TECHNOLOGIES 3182: 277-286. LECTURE NOTES IN COMPUTER SCIENCE
De Pasquale, JF; Meunier, JG. 2003. Categorisation techniques in computer-assisted reading and analysis of texts (CARAT) in the humanities. COMPUTERS AND THE HUMANITIES 37 (1): 111-118.
Debenham, J. 2004. Interacting with electronic institutions. DATABASE AND EXPERT SYSTEMS APPLICATIONS, PROCEEDINGS 3180: 181-190. LECTURE NOTES IN COMPUTER SCIENCE
Debenham, J; Simoff, S. 2005. Intelligent environments for next-generation e-markets. KNOWLEDGE-BASED INTELLIGENT INFORMATION AND ENGINEERING SYSTEMS, PT 1, PROCEEDINGS 3681: 751-757. LECTURE NOTES IN ARTIFICIAL INTELLIGENCE
Debnath, S; Mitra, P; Pal, N; Giles, CL. 2005. Automatic identification of informative sections of Web pages. IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING 17 (9): 1233-1246.
Debole, F; Sebastiani, F. 2005. An analysis of the relative hardness of Reuters-21578 subsets. JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE AND TECHNOLOGY 56 (6): 584-596.
DEERWESTER, S; DUMAIS, S; LANDAUER, T; FURNAS, G; BECK, L. 1988. IMPROVING INFORMATION-RETRIEVAL WITH LATENT SEMANTIC INDEXING. PROCEEDINGS OF THE ASIS ANNUAL MEETING 25: 36-40.
Degemmis, M; Lops, P; Ferilli, S; Di Mauro, N; Basile, TMA; Semeraro, G. 2005. Learning user profiles from text in e-commerce. ADVANCED DATA MINING AND APPLICATIONS, PROCEEDINGS 3584: 370-381. LECTURE NOTES IN ARTIFICIAL INTELLIGENCE

Delgado, M; Martin-Bautista, MJ; Sanchez, D; Serrano, JM; Vila, MA. 2002. Association rule extraction for text mining. FLEXIBLE QUERY ANSWERING SYSTEMS, PROCEEDINGS 2522: 154-162. LECTURE NOTES IN ARTIFICIAL INTELLIGENCE
Demaine, ED; Immorlica, N. 2003. Correlation clustering with partial information. APPROXIMATION, RANDOMIZATION, AND COMBINATORIAL OPTIMIZATION 2764: 1-13. LECTURE NOTES IN COMPUTER SCIENCE
Deng, ZH; Tang, SW. 2005. A non-VSM kNN algorithm for text classification. ADVANCED DATA MINING AND APPLICATIONS, PROCEEDINGS 3584: 339-346. LECTURE NOTES IN ARTIFICIAL INTELLIGENCE
Deng, ZH; Tang, SW; Yang, DQ; Zhang, M; Li, LY; Xie, KQ. 2004. A comparative study on feature weight in text categorization. ADVANCED WEB TECHNOLOGIES AND APPLICATIONS 3007: 588-597. LECTURE NOTES IN COMPUTER SCIENCE
Deng, ZH; Tang, SW; Yang, DQ; Zhang, M; Wu, XB; Yang, M. 2002. A linear text classification algorithm based on category relevance factors. DIGITAL LIBRARIES: PEOPLE, KNOWLEDGE, AND TECHNOLOGY, PROCEEDINGS 2555: 88-98. LECTURE NOTES IN COMPUTER SCIENCE
Deng, ZH; Tang, SW; Zhang, M. 2005. An efficient text categorization algorithm based on category memberships. FUZZY SYSTEMS AND KNOWLEDGE DISCOVERY, PT 1, PROCEEDINGS 3613: 374-382. LECTURE NOTES IN ARTIFICIAL INTELLIGENCE
Desai, M; Spink, A. 2005. An algorithm to cluster documents based on relevance. INFORMATION PROCESSING & MANAGEMENT 41 (5): 1035-1049.
Dhillon, IS; Modha, DS. 2000. A data-clustering algorithm on distributed memory multiprocessors. LARGE-SCALE PARALLEL DATA MINING 1759: 245-260. LECTURE NOTES IN ARTIFICIAL INTELLIGENCE
Dhillon, IS; Modha, DS. 2001. Concept decompositions for large sparse text data using clustering. MACHINE LEARNING 42 (1-2): 143-175.
Diao, YL; Lu, HJ; Wu, DK. 2000. A comparative study of classification based personal E-mail filtering. KNOWLEDGE DISCOVERY AND DATA MINING, PROCEEDINGS 1805: 408-419. LECTURE NOTES IN ARTIFICIAL INTELLIGENCE
Diaz, I; Ranilla, J; Montanes, E; Fernandez, J; Combarro, EF. 2004. Improving performance of text categorization by combining filtering and

support vector machines. JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE AND TECHNOLOGY 55 (7): 579-592.
Diederich, J; Kindermann, O; Leopold, E; Paass, G. 2003. Authorship attribution with support vector machines. APPLIED INTELLIGENCE 19 (1-2): 109-123.
DILLON, M; CAPLAN, P. 1980. TECHNIQUE FOR EVALUATING AUTOMATIC TERM CLUSTERING. JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE 31 (2): 89-96.
Ding, CHQ. 2005. A probabilistic model for Latent Semantic Indexing. JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE AND TECHNOLOGY 56 (6): 597-608.
Ding, J; Berleant, D. 2005. MedKit: a helper toolkit for automatic mining of MEDLINE/PubMed citations. BIOINFORMATICS 21 (5): 694-695.
Dobrokhotov, PB; Goutte, C; Veuthey, AL; Gaussier, E. 2005. Assisting medical annotation in Swiss-Prot using statistical classifiers. INTERNATIONAL JOURNAL OF MEDICAL INFORMATICS 74 (2-4): 317-324.
Dobrynin, V; Patterson, D; Rooney, N. 2004. Contextual document clustering. ADVANCES IN INFORMATION RETRIEVAL, PROCEEDINGS 2997: 167-180. LECTURE NOTES IN COMPUTER SCIENCE
Domedel-Puig, N; Wernisch, L. 2005. Applying GIFT, a gene interactions finder in text, to fly literature. BIOINFORMATICS 21 (17): 3582-3583.
Dorsey, RJ; Umhoefer, PJ; Falk, PD. 1997. Earthquake clustering inferred from Pliocene Gilbert-type fan deltas in the Loreto basin, Baja California Sur, Mexico. GEOLOGY 25 (8): 679-682.
Dransfield, E; Morrot, G; Martin, JF; Ngapo, TM. 2004. The application of a text clustering statistical analysis to aid the interpretation of focus group interviews. FOOD QUALITY AND PREFERENCE 15 (5): 477-488.
Dringus, LP; Ellis, T. 2005. Using data mining as a strategy for assessing asynchronous discussion forums. COMPUTERS & EDUCATION 45 (1): 141-160.
Dubhashi, D; Laura, L; Panconesi, A. 2003. Analysis and experimental evaluation of a simple algorithm for collaborative filtering in planted partition models. FST TCS 2003: FOUNDATIONS OF SOFTWARE TECHNOLOGY AND THEORETICAL COMPUTER SCIENCE 2914: 168-182. LECTURE NOTES IN COMPUTER SCIENCE
Dubois, V; Quafafou, M. 2002. Incremental and dynamic text mining - Graph structure discovery and visualization. FOUNDATIONS OF

INTELLIGENT SYSTEMS, PROCEEDINGS 2366: 265-273. LECTURE NOTES IN ARTIFICIAL INTELLIGENCE
Dumais, S. 2003. Data-driven approaches to information access. COGNITIVE SCIENCE 27 (3): 491-524.
Durbin, SD; Warner, D; Richter, HN; Gedeon, Z. 2002. Information self-service with a knowledge base that learns. AI MAGAZINE 23 (4): 41-49.
Efron, M. 2005. Eigenvalue-based model selection during latent semantic indexing. JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE AND TECHNOLOGY 56 (9): 969-988.
Eichmann, D; Srinivasan, P. 2002. Adaptive filtering of newswire stories using two-level clustering. INFORMATION RETRIEVAL 5 (2-3): 209-237.
Elmas, T; Ozkasap, O. 2004. Distributed document sharing with text classification over content-addressable network. CONTENT COMPUTING, PROCEEDINGS 3309: 70-81. LECTURE NOTES IN COMPUTER SCIENCE
Eom, JH; Zhang, BT. 2004. PubMiner: Machine learning-based text mining system for biomedical information mining. ARTIFICIAL INTELLIGENCE: METHODOLOGY, SYSTEMS, AND APPLICATIONS, PROCEEDINGS 3192: 216-225. LECTURE NOTES IN COMPUTER SCIENCE
Epshteyn, A; DeJong, G. 2005. Rotational prior knowledge for SVMs. MACHINE LEARNING: ECML 2005, PROCEEDINGS 3720: 108-119. LECTURE NOTES IN ARTIFICIAL INTELLIGENCE
Estabrooks, A; Jo, TH; Japkowicz, N. 2004. A multiple resampling method for learning from imbalanced data sets. COMPUTATIONAL INTELLIGENCE 20 (1): 18-36.
Esteban, AD. 2001. Integrating multilingual text classification tasks and user modeling in personalized newspaper services. USER MODELING 2001, PROCEEDINGS 2109: 268-270. LECTURE NOTES IN ARTIFICIAL INTELLIGENCE
Fan, KC; Wang, LS; Tu, YT. 1998. Classification of machine-printed and handwritten texts using character block layout variance. PATTERN RECOGNITION 31 (9): 1275-1284.
Fan, WG; Gordon, MD; Pathak, P. 2004. A generic ranking function discovery framework by genetic programming for information retrieval. INFORMATION PROCESSING & MANAGEMENT 40 (4): 587-602.
Fan, WG; Gordon, MD; Pathak, P. 2004. Discovery of context-specific ranking functions for effective information retrieval using genetic programming. IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING 16 (4): 523-527.

Fan, WG; Gordon, MD; Pathak, P. 2005. Genetic programming-based discovery of ranking functions for effective Web search. JOURNAL OF MANAGEMENT INFORMATION SYSTEMS 21 (4): 37-56.
Fan, XH; Sun, MS; Choi, K; Zhang, Q. 2005. Classifying Chinese texts in two steps. NATURAL LANGUAGE PROCESSING - IJCNLP 2005, PROCEEDINGS 3651: 302-313. LECTURE NOTES IN ARTIFICIAL INTELLIGENCE
Fatima, SS; Krishnan, R. 2001. Stylistic variation as a basis for genre-based text classification. IETE JOURNAL OF RESEARCH 47 (1-2): 59-63.
Feldman, R; Aumann, Y; Fresko, M; Liphstat, O; Rosenfeld, B; Schler, Y. 1999. Text mining via information extraction. PRINCIPLES OF DATA MINING AND KNOWLEDGE DISCOVERY 1704: 165-173. LECTURE NOTES IN ARTIFICIAL INTELLIGENCE
Feldman, R; Aumann, Y; Zilberstein, A; Ben-Yehuda, Y. 1998. Trend graphs: Visualizing the evolution of concept relationships in large document collections. PRINCIPLES OF DATA MINING AND KNOWLEDGE DISCOVERY 1510: 38-46. LECTURE NOTES IN ARTIFICIAL INTELLIGENCE
Feldman, R; Dagan, I; Hirsh, H. 1998. Mining text using keyword distributions. JOURNAL OF INTELLIGENT INFORMATION SYSTEMS 10 (3): 281-300.
Feldman, R; Fresko, M; Kinar, Y; Lindell, Y; Liphstat, O; Rajman, M; Schler, Y; Zamir, O. 1998. Text mining at the term level. PRINCIPLES OF DATA MINING AND KNOWLEDGE DISCOVERY 1510: 65-73. LECTURE NOTES IN ARTIFICIAL INTELLIGENCE
Feng, JL; Liu, HJ; Feng, YC. 2005. Sentential association based text classification systems. WEB TECHNOLOGIES RESEARCH AND DEVELOPMENT - APWEB 2005 3399: 1037-1040. LECTURE NOTES IN COMPUTER SCIENCE
Fernandez, J; Montanes, E; Diaz, I; Ranilla, J; Combarro, EF. 2004. Text categorization by a machine-learning-based term selection. DATABASE AND EXPERT SYSTEMS APPLICATIONS, PROCEEDINGS 3180: 253-262. LECTURE NOTES IN COMPUTER SCIENCE
Fierro, RD; Jiang, EP. 2005. Lanczos and the Riemannian SVD in information retrieval applications. NUMERICAL LINEAR ALGEBRA WITH APPLICATIONS 12 (4): 355-372.
Fine, S; Gilad-Bachrach, R; Shamir, E. 2002. Query by committee, linear separation and random walks. THEORETICAL COMPUTER SCIENCE 284 (1): 25-51.

Fisher, M; Everson, R. 2003. When are links useful? Experiments in text classification.. ADVANCES IN INFORMATION RETRIEVAL 2633: 41-56. LECTURE NOTES IN COMPUTER SCIENCE
Flesca, S; Manco, G; Masciari, E; Pontieri, L; Pugliese, A. 2005. Fast detection of XML structural similarity. IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING 17 (2): 160-175.
FOLTZ, PW; DUMAIS, ST. 1992. PERSONALIZED INFORMATION DELIVERY - AN ANALYSIS OF INFORMATION FILTERING METHODS. COMMUNICATIONS OF THE ACM 35 (12): 51-60.
Forman, G. 2005. Counting positives accurately despite inaccurate classification. MACHINE LEARNING: ECML 2005, PROCEEDINGS 3720: 564-575. LECTURE NOTES IN ARTIFICIAL INTELLIGENCE
Forman, G; Cohen, I. 2004. Learning from little: Comparison of classifiers given little training. KNOWLEDGE DISCOVERY IN DATABASES: PKDD 2004, PROCEEDINGS 3202: 161-172. LECTURE NOTES IN ARTIFICIAL INTELLIGENCE
Fragoudis, D; Meretakis, D; Likothanassis, S. 2005. Best terms: an efficient feature-selection algorithm for text categorization. KNOWLEDGE AND INFORMATION SYSTEMS 8 (1): 16-33.
Frasconi, P; Soda, G; Vullo, A. 2002. Hidden markov models for text categorization in multi-page documents. JOURNAL OF INTELLIGENT INFORMATION SYSTEMS 18 (2-3): 195-217.
Freeman, R; Yin, HJ. 2002. Self-organising maps for hierarchical tree view document clustering using contextual information. INTELLIGENT DATA ENGINEERING AND AUTOMATED LEARNING - IDEAL 2002 2412: 123-128. LECTURE NOTES IN COMPUTER SCIENCE
Freeman, RT; Yin, HJ. 2004. Adaptive topological tree structure for document organisation and visualisation. NEURAL NETWORKS 17 (8-9): 1255-1271.
Freeman, RT; Yin, HJ. 2005. Tree view self-organisation of web content. NEUROCOMPUTING 63: 415-446.
Freeman, RT; Yin, HJ. 2005. Web content management by self-organization. IEEE TRANSACTIONS ON NEURAL NETWORKS 16 (5): 1256-1268.
Friedman, C; Liu, HF; Shagina, L. 2003. A vocabulary development and visualization tool based on natural language processing and the mining of textual patient reports. JOURNAL OF BIOMEDICAL INFORMATICS 36 (3): 189-201.
Fu, P; Zhang, DY; Ma, ZF; Dong, H. 2005. SVM-based semantic text categorization for large scale web information organization. ADVANCES

IN NEURAL NETWORKS - ISNN 2005, PT 1, PROCEEDINGS 3496: 931-936. LECTURE NOTES IN COMPUTER SCIENCE
Fu, XH; Ma, ZF; Feng, BQ. 2004. Kernel-based semantic text categorization for large scale web information organization. GRID AND COOPERATIVE COMPUTING GCC 2004, PROCEEDINGS 3251: 389-396. LECTURE NOTES IN COMPUTER SCIENCE
Fujino, R; Arimura, H; Arikawa, S. 2000. Discovering unordered and ordered phrase association patterns for text mining. KNOWLEDGE DISCOVERY AND DATA MINING, PROCEEDINGS 1805: 281-293. LECTURE NOTES IN ARTIFICIAL INTELLIGENCE
Fuketa, M; Atlam, ES; Hanafusa, H; Morita, K; Kashiji, S; Mahmoud, R; Aoe, J. 2005. A new technique of determining speaker's intention for sentences in conversation. KNOWLEDGE-BASED INTELLIGENT INFORMATION AND ENGINEERING SYSTEMS, PT 4, PROCEEDINGS 3684: 612-618. LECTURE NOTES IN ARTIFICIAL INTELLIGENCE
Fuketa, M; Kadoya, Y; Atlam, E; Kunikata, T; Morita, K; Kashiji, S; Aoe, JI. 2005. A method of extracting and evaluating good and bad reputations for natural language expressions. INTERNATIONAL JOURNAL OF INFORMATION TECHNOLOGY & DECISION MAKING 4 (2): 177-196.
Fuketa, M; Lee, S; Tsuji, T; Okada, M; Aoe, J. 2000. A document classification method by using field association words. INFORMATION SCIENCES 126 (1-4): 57-70.
Fukuoka, K; Nakano, T; Inuzuka, N. 2005. Organising documents based on standard-example split test. KNOWLEDGE-BASED INTELLIGENT INFORMATION AND ENGINEERING SYSTEMS, PT 1, PROCEEDINGS 3681: 787-793. LECTURE NOTES IN ARTIFICIAL INTELLIGENCE
Furnkranz, J. 1999. Exploiting structural information for text classification on the WWW. ADVANCES IN INTELLIGENT DATA ANALYSIS, PROCEEDINGS 1642: 487-497. LECTURE NOTES IN COMPUTER SCIENCE
Galavotti, L; Sebastiani, F; Simi, M. 2000. Experiments on the use of feature selection and negative evidence in automated text categorization. RESEARCH AND ADVANCED TECHNOLOGY FOR DIGITAL LIBRARIES, PROCEEDINGS 1923: 59-68. LECTURE NOTES IN COMPUTER SCIENCE
Gao, B; Liu, TY; Feng, G; Qin, T; Cheng, QS; Ma, WY. 2005. Hierarchical taxonomy preparation for text categorization using consistent bipartite

spectral graph copartitioning. IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING 17 (9): 1263-1273.
Gao, J; Zhang, J. 2004. Text retrieval using sparsified concept decomposition matrix. COMPUTATIONAL AND INFORMATION SCIENCE, PROCEEDINGS 3314: 523-529. LECTURE NOTES IN COMPUTER SCIENCE
Gao, J; Zhang, J. 2005. Clustered SVD strategies in latent semantic indexing. INFORMATION PROCESSING & MANAGEMENT 41 (5): 1051-1063.
Gaussier, E; Goutte, C; Popat, K; Chen, F. 2002. A hierarchical model for clustering and categorising documents. ADVANCES IN INFORMATION REFTRIEVAL 2291: 229-247. LECTURE NOTES IN COMPUTER SCIENCE
Ge, Y; Li, XG; Bao, YB; Wang, DL. 2005. Evaluating document-to-document relevance based on document language model: Modeling, implementation and performance evaluation. COMPUTATIONAL LINGUISTICS AND INTELLIGENT TEXT PROCESSING 3406: 593-603. LECTURE NOTES IN COMPUTER SCIENCE
Gelbukh, A; Sidorov, G; Guzman-Arenas, A. 1999. Use of a weighted topic hierarchy for document classification. TEXT, SPEECH AND DIALOGUE 1692: 133-138. LECTURE NOTES IN ARTIFICIAL INTELLIGENCE
Geng, YB; Zhu, GM; Qiu, JR; Fan, JL; Zhang, JC. 2004. An experimental study of boosting model classifiers for Chinese text categorization. DIGITAL LIBRARIES: INTERNATIONAL COLLABORATION AND CROSS-FERTILIZATION, PROCEEDINGS 3334: 270-279. LECTURE NOTES IN COMPUTER SCIENCE
Gentili, GL; Marinilli, M; Micarelli, A; Sciarrone, F. 2001. Text categorization in an intelligent agent for filtering information on the Web. INTERNATIONAL JOURNAL OF PATTERN RECOGNITION AND ARTIFICIAL INTELLIGENCE 15 (3): 527-549.
Gerstenberger, MC; Wiemer, S; Jones, LM; Reasenberg, PA. 2005. Real-time forecasts of tomorrow's earthquakes in California. NATURE 435 (7040): 328-331.
Ghani, R. 2004. Mining the web to add semantics to retail data mining. WEB MINING: FROM WEB TO SEMANTIC WEB 3209: 43-56. LECTURE NOTES IN COMPUTER SCIENCE
Ghose, S; Jung, JG; Jo, GS. 2004. Collaborative detection of spam in peer-to-peer paradigm based on multi-agent systems. GRID AND COOPERATIVE COMPUTING GCC 2004, PROCEEDINGS 3251: 971-974. LECTURE NOTES IN COMPUTER SCIENCE

Gilardoni, F; Curcin, V; Karunanayake, K; Norgaard, J; Guo, Y. 2005. Integrated Informatics in life and materials sciences: An oxymoron?. QSAR & COMBINATORIAL SCIENCE 24 (1): 120-130.
Giorgetti, D; Sebastiani, F. 2003. Automating survey coding by multiclass text categorization techniques. JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE AND TECHNOLOGY 54 (14): 1269-1277.
Girgis, MR; Aly, AA. 2003. A feature selection and classification technique for text categorization. INTERNATIONAL JOURNAL OF COOPERATIVE INFORMATION SYSTEMS 12 (4): 441-454.
Glenisson, P; Coessens, B; Van Vooren, S; Mathys, J; Moreau, Y; De Moor, B. 2004. TXTGate: profiling gene groups with text-based information. GENOME BIOLOGY 5 (6): art. no.-R43.
Glenisson, P; Glanzel, W; Janssens, F; De Moor, B. 2005. Combining full text and bibliometric information in mapping scientific disciplines. INFORMATION PROCESSING & MANAGEMENT 41 (6): 1548-1572.
Godbole, S; Harpale, A; Sarawagi, S; Chakrabarti, S. 2004. Document classification through interactive supervision of document and term labels. KNOWLEDGE DISCOVERY IN DATABASES: PKDD 2004, PROCEEDINGS 3202: 185-196. LECTURE NOTES IN ARTIFICIAL INTELLIGENCE
Godbole, S; Harpale, A; Sarawagi, S; Chakrabarti, S. 2004. HIClass: Hyper-interactive text classification by interactive supervision of document and term labels. KNOWLEDGE DISCOVERY IN DATABASES: PKDD 2004, PROCEEDINGS 3202: 546-548. LECTURE NOTES IN ARTIFICIAL INTELLIGENCE
Godbole, S; Sarawagi, S. 2004. Discriminative methods for multi-labeled classification. ADVANCES IN KNOWLEDGE DISCOVERY AND DATA MINING, PROCEEDINGS 3056: 22-30. LECTURE NOTES IN ARTIFICIAL INTELLIGENCE
Goldman, JA; Chu, WW; Parker, DS; Goldman, RM. 1999. Term domain distribution analysis: A data mining tool for text databases. METHODS OF INFORMATION IN MEDICINE 38 (2): 96-101.
Gomez, JM; Cortizo, JC; Puertas, E; Ruiz, M. 2004. Concept indexing for automated text categorization. NATURAL LANGUAGE PROCESSING AND INFORMATION SYSTEMS 3136: 195-206. LECTURE NOTES IN COMPUTER SCIENCE
Goncalves, T; Quaresma, P. 2004. Using IR techniques to improve automated text classification. NATURAL LANGUAGE PROCESSING

AND INFORMATION SYSTEMS 3136: 374-379. LECTURE NOTES IN COMPUTER SCIENCE
GORDON, MD. 1991. USER-BASED DOCUMENT CLUSTERING BY REDESCRIBING SUBJECT DESCRIPTIONS WITH A GENETIC ALGORITHM. JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE 42 (5): 311-322.
Gordon, MD; Dumais, S. 1998. Using latent semantic indexing for literature based discovery. JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE 49 (8): 674-685.
Gordon, MD; Lindsay, RK. 1996. Toward discovery support systems: A replication, re-examination, and extension of Swanson's work on literature-based discovery of a connection between Raynaud's and fish oil. JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE 47 (2): 116-128.
Goren-Bar, D; Kuflik, T. 2005. Supporting user-subjective categorization with self-organizing maps and learning vector quantization. JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE AND TECHNOLOGY 56 (4): 345-355.
Goren-Bar, D; Kuflik, T; Lev, D; Shoval, P. 2001. Automating personal categorization using artificial neural networks. USER MODELING 2001, PROCEEDINGS 2109: 188-198. LECTURE NOTES IN ARTIFICIAL INTELLIGENCE
Greene, D; Cunningham, P. 2005. Producing accurate interpretable clusters from high-dimensional data. KNOWLEDGE DISCOVERY IN DATABASES: PKDD 2005 3721: 486-494. LECTURE NOTES IN ARTIFICIAL INTELLIGENCE
Guerrero-Bote, VP; Lopez-Pujalte, C; de Moya-Anegon, F; Herrero-Solana, V. 2003. Comparison of neural models for document clustering. INTERNATIONAL JOURNAL OF APPROXIMATE REASONING 34 (2-3): 287-305.
Gulla, JA; Brasethvik, T; Kaada, H. 2004. A flexible workbench for document analysis and text mining. NATURAL LANGUAGE PROCESSING AND INFORMATION SYSTEMS 3136: 336-347. LECTURE NOTES IN COMPUTER SCIENCE
Guo, D; Berry, MW; Thompson, BB; Bailin, S. 2003. Knowledge-enhanced latent semantic indexing. INFORMATION RETRIEVAL 6 (2): 225-250.
Guo, GD; Wang, H; Bell, D; Bi, YX; Greer, K. 2004. An kNN model-based approach and its application in text categorization. COMPUTATIONAL LINGUISTICS AND INTELLIGENT TEXT PROCESSING 2945: 559-570. LECTURE NOTES IN COMPUTER SCIENCE

Guthrie, L; Liu, W; Xia, YQ. 2005. Text classification with tournament methods. TEXT, SPEECH AND DIALOGUE, PROCEEDINGS 3658: 77-84. LECTURE NOTES IN ARTIFICIAL INTELLIGENCE
Gutwin, C; Paynter, G; Witten, I; Nevill-Manning, C; Frank, E. 1999. Improving browsing in digital libraries with keyphrase indexes. DECISION SUPPORT SYSTEMS 27 (1-2): 81-104.
Gweon, G; Rose, CP; Wittwer, J; Nueckles, M. 2005. Supporting efficient and reliable content analysis using automatic text processing technology. HUMAN-COMPUTER INTERACTION - INTERACT 2005, PROCEEDINGS 3585: 1112-1115. LECTURE NOTES IN COMPUTER SCIENCE
Haddad, H. 2003. French noun phrase indexing and mining for an information retrieval system. STRING PROCESSING AND INFORMATION RETRIEVAL, PROCEEDINGS 2857: 277-286. LECTURE NOTES IN COMPUTER SCIENCE
Hadjarian, A; Bala, J; Pachowicz, P. 2001. Text categorization through multistrategy learning and visualization. COMPUTATIONAL LINGUISTICS AND INTELLIGENT TEXT PROCESSING 2004: 437-443. LECTURE NOTES IN COMPUTER SCIENCE
Hakenberg, J; Schmeier, S; Kowald, A; Klipp, E; Leser, U. 2004. Finding kinetic parameters using text mining. OMICS-A JOURNAL OF INTEGRATIVE BIOLOGY 8 (2): 131-152.
Halkidi, M; Nguyen, B; Varlamis, I; Vazirgiannis, M. 2003. THESUS: Organizing Web document collections based on link semantics. VLDB JOURNAL 12 (4): 320-332.
Hamilton, RJ. 2004. Material appropriate processing and elaboration: The impact of balanced and complementary types of processing on learning concepts from text. BRITISH JOURNAL OF EDUCATIONAL PSYCHOLOGY 74: 221-237, Part 2.
Hammouda, KM; Kamel, MS. 2004. Efficient phrase-based document indexing for web document clustering. IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING 16 (10): 1279-1296.
Hammouda, KM; Matute, DN; Kamel, MS. 2005. CorePhrase: Keyphrase extraction for document clustering. MACHINE LEARNING AND DATA MINING IN PATTERN RECOGNITION, PROCEEDINGS 3587: 265-274. LECTURE NOTES IN ARTIFICIAL INTELLIGENCE
Han, L; Yuan, XB; Tang, QY; Kustra, R. 2004. An efficient method to estimate labelled sample size for transductive LDA(QDA/MDA) based on Bayes risk. MACHINE LEARNING: ECML 2004, PROCEEDINGS 3201: 274-285. LECTURE NOTES IN COMPUTER SCIENCE

Han, XX; Zu, GW; Ohyama, W; Wakabayashi, T; Kimura, F. 2004. Accuracy improvement of automatic text classification based on feature transformation and multi-classifier combination. CONTENT COMPUTING, PROCEEDINGS 3309: 463-468. LECTURE NOTES IN COMPUTER SCIENCE
Hare, JS; Lewis, PH. 2005. On image retrieval using salient regions with vector-spaces and latent semantics. IMAGE AND VIDEO RETRIEVAL, PROCEEDINGS 3568: 540-549. LECTURE NOTES IN COMPUTER SCIENCE
Hatzivassiloglou, V; Weng, WB. 2002. Learning anchor verbs for biological interaction patterns from published text articles. INTERNATIONAL JOURNAL OF MEDICAL INFORMATICS 67 (1-3): 19-32.
He, J; Tan, AH; Tan, CL. 2003. On machine learning methods for Chinese document categorization. APPLIED INTELLIGENCE 18 (3): 311-322.
He, Q. 1999. Knowledge discovery through co-word analysis. LIBRARY TRENDS 48 (1): 133-159.
He, QM; Qiu, L; Zhao, GT; Wang, SK. 2004. Text categorization based on domain ontology. WEB INFORMATION SYSTEMS - WISE 2004, PROCEEDINGS 3306: 319-324. LECTURE NOTES IN COMPUTER SCIENCE
He, X; Zha, HY; Ding, CHQ; Simon, HD. 2002. Web document clustering using hyperlink structures. COMPUTATIONAL STATISTICS & DATA ANALYSIS 41 (1): 19-45.
He, Y; Hui, SC; Fong, ACM. 2002. Mining a Web citation database for document clustering. APPLIED ARTIFICIAL INTELLIGENCE 16 (4): 283-302.
Heeren, F; Sihn, W. 2002. Message analysis for the recommendation of contact persons within defined subject fields. DEVELOPMENTS IN APPLIED ARTIFICAIL INTELLIGENCE, PROCEEDINGS 2358: 660-669. LECTURE NOTES IN ARTIFICIAL INTELLIGENCE
Hein, M; Lal, TN; Bousquet, O. 2004. Hilbertian metrics on probability measures and their application in SVM's. PATTERN RECOGNITION 3175: 270-277. LECTURE NOTES IN COMPUTER SCIENCE
Heinze, DT; Morsch, ML; Holbrook, J. 2001. Mining free-text medical records. JOURNAL OF THE AMERICAN MEDICAL INFORMATICS ASSOCIATION: 254-258, Suppl. S.
Helmstetter, A; Kagan, YY; Jackson, DD. 2005. Importance of small earthquakes for stress transfers and earthquake triggering. JOURNAL OF GEOPHYSICAL RESEARCH-SOLID EARTH 110 (B5): art. no.-B05S08.

Hidalgo, JMG; Garcia, FC; Sanz, EP. 2005. Named entity recognition for Web content filtering. NATURAL LANGUAGE PROCESSING AND INFORMATION SYSTEMS, PROCEEDINGS 3513: 286-297. LECTURE NOTES IN COMPUTER SCIENCE
Hidalgo, JMG; Rodriguez, MD; Perez, JCC. 2005. The role of word sense disambiguation in automated text categorization. NATURAL LANGUAGE PROCESSING AND INFORMATION SYSTEMS, PROCEEDINGS 3513: 298-309. LECTURE NOTES IN COMPUTER SCIENCE
Higuchi, K. 2004. Computer assisted quantitative analysis of newspaper articles. SOCIOLOGICAL THEORY AND METHODS 19 (2): 161-176.
Higuchi, K. 2004. Quantitative analysis of textual data: Differentiation and coordination of two approaches. SOCIOLOGICAL THEORY AND METHODS 19 (1): 101-115.
Hirsch, L; Saeedi, M; Hirsch, R. 2005. Evolving rules for document classification. GENETIC PROGRAMMING, PROCEEDINGS 3447: 85-95. LECTURE NOTES IN COMPUTER SCIENCE
Hirsch, L; Saeedi, M; Hirsch, R. 2005. Evolving text classification rules with genetic programming. APPLIED ARTIFICIAL INTELLIGENCE 19 (7): 659-676.
Ho, TB; Nguyen, NB. 2002. Nonhierarchical document clustering based on a tolerance rough set model. INTERNATIONAL JOURNAL OF INTELLIGENT SYSTEMS 17 (2): 199-212.
Hoenkamp, E. 2003. Unitary operators on the document space. JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE AND TECHNOLOGY 54 (4): 314-320.
Homayouni, R; Heinrich, K; Wei, L; Berry, MW. 2005. Gene clustering by Latent Semantic Indexing of MEDLINE abstracts. BIOINFORMATICS 21 (1): 104-115.
Hong, SJ; Weiss, SM. 2001. Advances in predictive models for data mining. PATTERN RECOGNITION LETTERS 22 (1): 55-61, Sp. Iss. SI.
Horng, JT; Yeh, CC. 2000. Applying genetic algorithms to query optimization in document retrieval. INFORMATION PROCESSING & MANAGEMENT 36 (5): 737-759.
Horng, YJ; Chen, SM; Chang, YC; Lee, CH. 2005. A new method for fuzzy information retrieval based on fuzzy hierarchical clustering and fuzzy inference techniques. IEEE TRANSACTIONS ON FUZZY SYSTEMS 13 (2): 216-228.
Hotho, A; Maedche, A; Staab, S; Studer, R. 2001. SEAL-II - The soft spot between richly structured and unstructured knowledge. JOURNAL OF UNIVERSAL COMPUTER SCIENCE 7 (7): 566-590.

Hotho, A; Staab, S; Stumme, G. 2003. Explaining text clustering results using semantic structures. KNOWLEDGE DISCOVERY IN DATABASES: PKDD 2003, PROCEEDINGS 2838: 217-228. LECTURE NOTES IN ARTIFICIAL INTELLIGENCE
How, BC; Kiong, WT. 2005. An examination of feature selection frameworks in text categorization. INFORMATION RETRIEVAL TECHNOLOGY, PROCEEDINGS 3689: 558-564. LECTURE NOTES IN COMPUTER SCIENCE
Howland, P; Jeon, M; Park, H. 2003. Structure preserving dimension reduction for clustered text data based on the generalized singular value decomposition. SIAM JOURNAL ON MATRIX ANALYSIS AND APPLICATIONS 25 (1): 165-179.
Howland, P; Park, H. 2004. Generalizing discriminant analysis using the generalized singular value decomposition. IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE 26 (8): 995-1006.
Hristovski, D; Peterlin, B; Mitchell, JA; Humphrey, SM. 2005. Using literature-based discovery to identify disease candidate genes. INTERNATIONAL JOURNAL OF MEDICAL INFORMATICS 74 (2-4): 289-298.
Hu, Y; Duan, JY; Chen, XM; Pei, BZ; Lu, RZ. 2005. A new method for sentiment classification in text retrieval. NATURAL LANGUAGE PROCESSING - IJCNLP 2005, PROCEEDINGS 3651: 1-9. LECTURE NOTES IN ARTIFICIAL INTELLIGENCE
Hu, YH; Hines, LM; Weng, HF; Zuo, DM; Rivera, M; Richardson, A; LaBaer, J. 2003. Analysis of genomic and proteomic data using advanced literature mining. JOURNAL OF PROTEOME RESEARCH 2 (4): 405-412.
Hu, ZZ; Mani, I; Hermoso, V; Liu, HF; Wu, CH. 2004. IProLINK: an integrated protein resource for literature mining. COMPUTATIONAL BIOLOGY AND CHEMISTRY 28 (5-6): 409-416.
Huang, CC; Chuang, SL; Chien, LF. 2005. Categorizing unknown text segments for information extraction using a search result mining approach. NATURAL LANGUAGE PROCESSING - IJCNLP 2004 3248: 576-586. LECTURE NOTES IN COMPUTER SCIENCE
Huang, S; Xue, GR; Zhang, BY; Chen, Z; Yu, Y; Ma, WY. 2004. Multi-type features based Web document clustering. WEB INFORMATION SYSTEMS - WISE 2004, PROCEEDINGS 3306: 253-265. LECTURE NOTES IN COMPUTER SCIENCE
Huang, TM; Kecman, V. 2004. Semi-supervised learning from unbalanced labeled data - An improvement. KNOWLEDGE-BASED INTELLIGENT

INFORMATION AND ENGINEERING SYSTEMS, PT 3, PROCEEDINGS 3215: 802-808. LECTURE NOTES IN ARTIFICIAL INTELLIGENCE
Huang, W; Nakamori, Y; Wang, SY; Ma, TJ. 2004. Mining medline for new possible relations of concepts. COMPUTATIONAL AND INFORMATION SCIENCE, PROCEEDINGS 3314: 794-799. LECTURE NOTES IN COMPUTER SCIENCE
Huang, XC; Chen, J; Yan, PL; Luo, X. 2005. Word segmentation and POS tagging for Chinese keyphrase extraction. ADVANCED DATA MINING AND APPLICATIONS, PROCEEDINGS 3584: 364-369. LECTURE NOTES IN ARTIFICIAL INTELLIGENCE
Huang, XC; Wu, M; Xia, DL; Yan, PL. 2005. Difference-similitude matrix in text classification. FUZZY SYSTEMS AND KNOWLEDGE DISCOVERY, PT 2, PROCEEDINGS 3614: 21-30, Part 2. LECTURE NOTES IN ARTIFICIAL INTELLIGENCE
Hubert, G. 2005. A voting method for XML retrieval. ADVANCES IN XML INFORMATION RETRIEVAL 3493: 183-195. LECTURE NOTES IN COMPUTER SCIENCE
Hudomalj, E; Vidmar, G. 2003. OLAP and bibliographic databases. SCIENTOMETRICS 58 (3): 609-622.
Hung, CL; Wermter, S; Smith, P. 2004. Hybrid neural document clustering using guided self-organization and wordnet. IEEE INTELLIGENT SYSTEMS 19 (2): 68-77.
Hung, CL; Wermter, S; Smith, P. 2004. Predictive top-down knowledge improves neural exploratory bottom-up clustering. ADVANCES IN INFORMATION RETRIEVAL, PROCEEDINGS 2997: 154-166. LECTURE NOTES IN COMPUTER SCIENCE
Hung, CM; Chien, LF. 2005. Text classification using Web corpora and EM algorithms. INFORMATION RETRIEVAL TECHNOLOGY 3411: 12-23. LECTURE NOTES IN COMPUTER SCIENCE
Husbands, P; Simon, H; Ding, C. 2005. Term norm distribution and its effects on Latent Semantic Indexing. INFORMATION PROCESSING & MANAGEMENT 41 (4): 777-787.
Hussin, MF; Kamel, MS; Nagi, MH. 2004. An efficient two-level SOMART document clustering through dimensionality reduction. NEURAL INFORMATION PROCESSING 3316: 158-165. LECTURE NOTES IN COMPUTER SCIENCE
Hwang, BY; Lee, BJ. 2004. An efficient e-mail monitoring system for detecting proprietary information outflow using broad concept learning.

METAINFORMATICS 3002: 72-78. LECTURE NOTES IN COMPUTER SCIENCE
Hwang, JH; Ryu, KH. 2004. A new XML clustering for structural retrieval. CONCEPTUAL MODELING - ER 2004, PROCEEDINGS 3288: 377-387. LECTURE NOTES IN COMPUTER SCIENCE
Hwang, JH; Ryu, KH. 2005. A new sequential mining approach to XML document clustering. WEB TECHNOLOGIES RESEARCH AND DEVELOPMENT - APWEB 2005 3399: 266-276. LECTURE NOTES IN COMPUTER SCIENCE
Hwang, JH; Ryu, KH. 2005. Clustering and retrieval of XML documents by structure. COMPUTATIONAL SCIENCE AND ITS APPLICATIONS - ICCSA 2005, PT 2 3480: 925-935. LECTURE NOTES IN COMPUTER SCIENCE
Im, Y; Song, J; Park, D. 2005. Fuzzy post-clustering algorithm for web search engine. INFORMATION RETRIEVAL TECHNOLOGY, PROCEEDINGS 3689: 709-714. LECTURE NOTES IN COMPUTER SCIENCE
Inoue, K; Urahama, K. 2001. Fuzzy clustering based on cooccurrence matrix and its application to data retrieval. ELECTRONICS AND COMMUNICATIONS IN JAPAN PART II-ELECTRONICS 84 (8): 10-19.
Itert, L; Duch, W; Pestian, J. 2005. Medical document categorization using a Priori Knowledge. ARTIFICIAL NEURAL NETWORKS: BIOLOGICAL INSPIRATIONS - ICANN 2005, PT 1, PROCEEDINGS 3696: 641-646. LECTURE NOTES IN COMPUTER SCIENCE
JACOBS, PS; RAU, LF. 1993. INNOVATIONS IN TEXT INTERPRETATION. ARTIFICIAL INTELLIGENCE 63 (1-2): 143-191.
Jacquenet, FO; Largeron, C. 2004. Discovering unexpected information for technology watch. KNOWLEDGE DISCOVERY IN DATABASES: PKDD 2004, PROCEEDINGS 3202: 219-230. LECTURE NOTES IN ARTIFICIAL INTELLIGENCE
Jebara, T; Kondor, R. 2003. Bhattacharyya and expected likelihood kernels. LEARNING THEORY AND KERNEL MACHINES 2777: 57-71. LECTURE NOTES IN ARTIFICIAL INTELLIGENCE
Jeong, OR; Cho, DS. 2004. A personalized recommendation agent system for e-mail document classification. COMPUTATIONAL SCIENCE AND ITS APPLICATIONS - ICCSA 2004, PT 3 3045: 558-565. LECTURE NOTES IN COMPUTER SCIENCE
Jiang, EP; Berry, MW. 2000. Solving total least-squares problems in information retrieval. LINEAR ALGEBRA AND ITS APPLICATIONS 316 (1-3): 137-156.

Jing, LP; Ng, MK; Xu, J; Huang, JZ. 2005. Subspace clustering of text documents with feature weighting K-means algorithm. ADVANCES IN KNOWLEDGE DISCOVERY AND DATA MINING, PROCEEDINGS 3518: 802-812. LECTURE NOTES IN ARTIFICIAL INTELLIGENCE
Jing, LP; Ng, MK; Xu, J; Huang, JZX. 2005. On the performance of feature weighting K-means for text subspace clustering. ADVANCES IN WEB-AGE INFORMATION MANAGEMENT, PROCEEDINGS 3739: 502-512. LECTURE NOTES IN COMPUTER SCIENCE
Johnson, DE; Oles, FJ; Zhang, T; Goetz, T. 2002. A decision-tree-based symbolic rule induction system for text categorization. IBM SYSTEMS JOURNAL 41 (3): 428-437.
Jones, S; Paynter, GW. 2002. Automatic extraction of document keyphrases for use in digital libraries: Evaluation and applications. JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE AND TECHNOLOGY 53 (8): 653-677.
Joo, KH; Lee, WS. 2005. An incremental document clustering for the large document database. INFORMATION RETRIEVAL TECHNOLOGY, PROCEEDINGS 3689: 374-387. LECTURE NOTES IN COMPUTER SCIENCE
Jorgensen, P. 2005. Incorporating context in text analysis by interactive activation with competition artificial neural networks. INFORMATION PROCESSING & MANAGEMENT 41 (5): 1081-1099.
Juan, A; Vidal, E. 2002. On the use of Bernoulli mixture models for text classification. PATTERN RECOGNITION 35 (12): 2705-2710.
Jung, SY; Hong, JH; Kim, TS. 2005. A statistical model for user preference. IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING 17 (6): 834-843.
Kaban, A; Girolami, MA. 2002. Fast extraction of semantic features from a latent semantic indexed text corpus. NEURAL PROCESSING LETTERS 15 (1): 31-34.
Kadoya, Y; Atlam, ES; Morita, K; Fuketa, M; Sumitomo, T; Aoe, J. 2004. A new classification method of determining the speaker's intention for sentences in conversation. KNOWLEDGE-BASED INTELLIGENT INFORMATION AND ENGINEERING SYSTEMS, PT 1, PROCEEDINGS 3213: 549-557. LECTURE NOTES IN COMPUTER SCIENCE
Kadoya, Y; Morita, K; Fuketa, M; Oono, M; Atlam, ES; Sumitomo, T; Aoe, JI. 2005. A sentence classification technique using intention association expressions. INTERNATIONAL JOURNAL OF COMPUTER MATHEMATICS 82 (7): 777-792.

KAGAN, YY; JACKSON, DD. 1991. LONG-TERM EARTHQUAKE CLUSTERING. GEOPHYSICAL JOURNAL INTERNATIONAL 104 (1): 117-133.
Kaki, M; Aula, A. 2005. Findex: improving search result use through automatic filtering categories. INTERACTING WITH COMPUTERS 17 (2): 187-206.
Kando, N. 2003. CLIR at NTCIR workshop 3: Cross-language and cross-genre retrieval. ADVANCES IN CROSS-LANGUAGE INFORMATION RETRIEVAL 2785: 485-504. LECTURE NOTES IN COMPUTER SCIENCE
Kando, N. 2003. Evaluation of information access technologies at the NTCIR workshop. COMPARATIVE EVALUATION OF MULTILINGUAL INFORMATION ACCESS SYSTEMS 3237: 29-43. LECTURE NOTES IN COMPUTER SCIENCE
Kaneda, Y; Ueda, N; Saito, K. 2004. Extended parametric mixture model for robust multi-labeled text categorization. KNOWLEDGE-BASED INTELLIGENT INFORMATION AND ENGINEERING SYSTEMS, PT 2, PROCEEDINGS 3214: 616-623. LECTURE NOTES IN COMPUTER SCIENCE
Kang, DK; Zhang, J; Silvescu, A; Honavar, V. 2005. Multinomial event model based abstraction for sequence and text classification. ABSTRACTION, REFORMULATION AND APPROXIMATION, PROCEEDINGS 3607: 134-148. LECTURE NOTES IN ARTIFICIAL INTELLIGENCE
Kang, JH; Ryu, KR; Kwon, HC. 2004. Using cluster-based sampling to select initial training set for active learning in text classification. ADVANCES IN KNOWLEDGE DISCOVERY AND DATA MINING, PROCEEDINGS 3056: 384-388. LECTURE NOTES IN ARTIFICIAL INTELLIGENCE
Kang, SS. 2004. Term-specific language modeling approach to text categorization. COMPUTATIONAL SCIENCE AND ITS APPLICATIONS - ICCSA 2004, PT 4 3046: 735-742. LECTURE NOTES IN COMPUTER SCIENCE
Kang, YH. 2005. Representative terrn based feature selection method for SVM based document classification. KNOWLEDGE-BASED INTELLIGENT INFORMATION AND ENGINEERING SYSTEMS, PT 1, PROCEEDINGS 3681: 56-61. LECTURE NOTES IN ARTIFICIAL INTELLIGENCE
Karopka, T; Scheel, T; Bansemer, S; Glass, A. 2004. Automatic construction of gene relation networks using text mining and gene expression data.

MEDICAL INFORMATICS AND THE INTERNET IN MEDICINE 29 (2): 169-183.
Karras, DA; Mertzios, BG. 2002. A robust meaning extraction methodology using supervised neural networks. AL 2002: ADVANCES IN ARTIFICIAL INTELLIGENCE 2557: 498-510. LECTURE NOTES IN ARTIFICIAL INTELLIGENCE
Katoh, M; Katoh, M. 2004. Characterization of FMN2 gene at human chromosome 1q43. INTERNATIONAL JOURNAL OF MOLECULAR MEDICINE 14 (3): 469-474.
Katoh, M; Katoh, M. 2004. Identification and characterization of human HESL, rat Hesl and rainbow trout hesl genes in silico. INTERNATIONAL JOURNAL OF MOLECULAR MEDICINE 14 (4): 747-751.
Katoh, M; Katoh, M. 2004. Identification and characterization of human DFNA5L, mouse DFNA5L, and rat DFNA5L genes in silico. INTERNATIONAL JOURNAL OF ONCOLOGY 25 (3): 765-770.
Katoh, M; Katoh, M. 2004. Identification and characterization of human CKTSF1B2 and CKTSF1B3 genes in silico. ONCOLOGY REPORTS 12 (2): 423-427.
Katoh, M; Katoh, M. 2004. Identification and characterization of human HES2, HES3, and HES5 genes in silico. INTERNATIONAL JOURNAL OF ONCOLOGY 25 (2): 529-534.
Katoh, M; Katoh, M. 2004. Identification and characterization of human ARHGAP23 gene in silico. INTERNATIONAL JOURNAL OF ONCOLOGY 25 (2): 535-540.
Katoh, M; Katoh, M. 2004. Identification and characterization of human FCHO2 and mouse Fcho2 genes in silico. INTERNATIONAL JOURNAL OF MOLECULAR MEDICINE 14 (2): 327-331.
Katoh, M; Katoh, M. 2004. Identification and characterization of ARHGAP24 and ARHGAP25 genes in silico. INTERNATIONAL JOURNAL OF MOLECULAR MEDICINE 14 (2): 333-338.
Katoh, M; Katoh, M. 2004. Identification and characterization of the human FMN1 gene in silico. INTERNATIONAL JOURNAL OF MOLECULAR MEDICINE 14 (1): 121-126.
Katoh, M; Katoh, M. 2004. Identification and characterization of human FOXK1 gene in silico. INTERNATIONAL JOURNAL OF MOLECULAR MEDICINE 14 (1): 127-132.
Katoh, M; Katoh, M. 2005. Comparative genomics on nemo-like kinase gene. INTERNATIONAL JOURNAL OF ONCOLOGY 26 (6): 1715-1719.

Katoh, M; Katoh, M. 2005. Identification and characterization of rat Ankrd6 gene in silico. INTERNATIONAL JOURNAL OF MOLECULAR MEDICINE 15 (2): 359-363.
Katoh, Y; Katoh, M. 2004. Identification and characterization of CDC50A, CDC50B and CDC50C genes in silico. ONCOLOGY REPORTS 12 (4): 939-943.
Katoh, Y; Katoh, M. 2004. KIF27 is one of orthologs for Drosophila Costal-2. INTERNATIONAL JOURNAL OF ONCOLOGY 25 (6): 1875-1880.
Kawahara, M; Kawano, H. 1999. ROC performance evaluation of web-based bibliographic navigator using extended association rules. INTERNET APPLICATIONS 1749: 216-225. LECTURE NOTES IN COMPUTER SCIENCE
Kawano, H; Kawahara, M. 2000. Mondou: Information navigator with visual interface. DATA WAREHOUSING AND KNOWLEDGE DISCOVERY, PROCEEDINGS 1874: 425-430. LECTURE NOTES IN COMPUTER SCIENCE
Kawano, M; Watada, J; Kawaura, T. 2005. Data mining method from text databases. KNOWLEDGE-BASED INTELLIGENT INFORMATION AND ENGINEERING SYSTEMS, PT 3, PROCEEDINGS 3683: 1122-1128. LECTURE NOTES IN ARTIFICIAL INTELLIGENCE
Kayed, A; Colomb, RM. 2002. Extracting ontological concepts for tendering conceptual structures. DATA & KNOWLEDGE ENGINEERING 40 (1): 71-89.
Kazama, J; Tsujii, J. 2005. Maximum entropy models with inequality constraints: A case study on text categorization. MACHINE LEARNING 60 (1-3): 159-194.
Kehagias, A; Petridis, V; Kaburlasos, VG; Fragkou, P. 2003. A comparison of word- and sense-based text categorization using several classification algorithms. JOURNAL OF INTELLIGENT INFORMATION SYSTEMS 21 (3): 227-247.
Keller, M; Bengio, S. 2005. A neural network for text representation. ARTIFICIAL NEURAL NETWORKS: FORMAL MODELS AND THEIR APPLICATIONS - ICANN 2005, PT 2, PROCEEDINGS 3697: 667-672. LECTURE NOTES IN COMPUTER SCIENCE
Khan, MS; Khor, SW. 2004. Web document clustering using a hybrid neural network. APPLIED SOFT COMPUTING 4 (4): 423-432.
Kibriya, AM; Frank, E; Pfahringer, B; Holmes, G. 2004. Multinomial naive Bayes for text categorization revisited. AI 2004: ADVANCES IN ARTIFICIAL INTELLIGENCE, PROCEEDINGS 3339: 488-499. LECTURE NOTES IN ARTIFICIAL INTELLIGENCE

Kim, BM; Li, Q; Lee, KH; Kang, BY. 2005. Extraction of representative keywords considering co-occurrence in positive documents. FUZZY SYSTEMS AND KNOWLEDGE DISCOVERY, PT 2, PROCEEDINGS 3614: 752-761, Part 2. LECTURE NOTES IN ARTIFICIAL INTELLIGENCE
Kim, D; Jung, HM; Lee, GG. 2003. Unsupervised learning of mDTD extraction patterns for Web text mining. INFORMATION PROCESSING & MANAGEMENT 39 (4): 623-637.
Kim, HJ; Kim, J. 2004. Combining active learning and boosting for Naive Bayes text classifiers. ADVANCES IN WEB-AGE INFORMATION MANAGEMENT: PROCEEDINGS 3129: 519-527. LECTURE NOTES IN COMPUTER SCIENCE
Kim, HJ; Kim, JU; Ra, YG. 2005. Boosting Naive Bayes text classification using uncertainty-based selective sampling. NEUROCOMPUTING 67: 403-410.
Kim, HJ; Lee, SG. 2002. User feedback-driven document clustering technique for information organization. IEICE TRANSACTIONS ON INFORMATION AND SYSTEMS E85D (6): 1043-1048.
Kim, HJ; Lee, SG. 2004. An intelligent information system for organizing online text documents. KNOWLEDGE AND INFORMATION SYSTEMS 6 (2): 125-149.
Kim, J; Kim, MH. 2004. An evaluation of passage-based text categorization. JOURNAL OF INTELLIGENT INFORMATION SYSTEMS 23 (1): 47-65.
Kim, JH. 2002. Bioinformatics and genomic medicine. GENETICS IN MEDICINE 4 (6): 62S-65S, Suppl. S.
Kim, LC; Myoung, SH. 2003. Text categorization using hybrid multiple model schemes. ADVANCES IN INTELLIGENT DATA ANALYSIS V 2810: 88-99. LECTURE NOTES IN COMPUTER SCIENCE
Kim, S; Fox, EA. 2004. Interest-based user grouping model for collaborative filtering in digital libraries. DIGITAL LIBRARIES: INTERNATIONAL COLLABORATION AND CROSS-FERTILIZATION, PROCEEDINGS 3334: 533-542. LECTURE NOTES IN COMPUTER SCIENCE
Kim, SB; Rim, HC. 2004. Recomputation of class relevance scores for improving text classification. COMPUTATIONAL LINGUISTICS AND INTELLIGENT TEXT PROCESSING 2945: 580-583. LECTURE NOTES IN COMPUTER SCIENCE
Kim, SB; Rim, HC; Kim, JD. 2005. Topic document model approach for naive Bayes text classification. IEICE TRANSACTIONS ON INFORMATION AND SYSTEMS E88D (5): 1091-1094.

Kim, SW; Oommen, BJ. 2002. Optimizing Kernel-based Nonlinear Subspace methods using Prototype Reduction Schemes. AL 2002: ADVANCES IN ARTIFICIAL INTELLIGENCE 2557: 155-166. LECTURE NOTES IN ARTIFICIAL INTELLIGENCE
Kim, SW; Oommen, BJ. 2004. Enhancing prototype reduction schemes with recursion: A method applicable for "large" data sets. IEEE TRANSACTIONS ON SYSTEMS MAN AND CYBERNETICS PART B-CYBERNETICS 34 (3): 1384-1397.
Kim, SW; Oommen, BJ. 2004. On using prototype reduction schemes to optimize kernel-based nonlinear subspace methods. PATTERN RECOGNITION 37 (2): 227-239.
Kim, YS; Chang, JH; Zhang, BT. 2003. An empirical study on dimensionality optimization in text mining for linguistic knowledge acquisition. ADVANCES IN KNOWLEDGE DISCOVERY AND DATA MINING 2637: 111-116. LECTURE NOTES IN ARTIFICIAL INTELLIGENCE
Kise, K; Junker, M; Dengel, A; Matsumoto, K. 2003. Effectiveness of passage-based document retrieval for short queries. IEICE TRANSACTIONS ON INFORMATION AND SYSTEMS E86D (9): 1753-1761.
Kleinberg, J. 2003. Bursty and hierarchical structure in streams. DATA MINING AND KNOWLEDGE DISCOVERY 7 (4): 373-397.
Klopotek, MA. 2002. A new Bayesian tree learning method with reduced time and space complexity. FUNDAMENTA INFORMATICAE 49 (4): 349-367.
Klopotek, MA. 2002. Mining Bayesian network structure for large sets of variables. FOUNDATIONS OF INTELLIGENT SYSTEMS, PROCEEDINGS 2366: 114-122. LECTURE NOTES IN ARTIFICIAL INTELLIGENCE
Klopotek, MA. 2005. Very large Bayesian multinets for text classification. FUTURE GENERATION COMPUTER SYSTEMS 21 (7): 1068-1082.
Klopotek, MA; Wierzchon, ST; Ciesielski, K; Draminski, M; Czerski, D. 2005. Coexistence of fuzzy and crisp concepts in document maps. ARTIFICIAL NEURAL NETWORKS: FORMAL MODELS AND THEIR APPLICATIONS - ICANN 2005, PT 2, PROCEEDINGS 3697: 859-864. LECTURE NOTES IN COMPUTER SCIENCE
Klopotek, MA; Woch, M. 2003. Very large Bayesian networks in text classification. COMPUTATIONAL SCIENCE - ICCS 2003, PT I, PROCEEDINGS 2657: 397-406. LECTURE NOTES IN COMPUTER SCIENCE

Ko, HM; Lam, W. 2005. A new approach for semi-supervised online news classification. WEB AND COMMUNICATION TECHNOLOGIES AND INTERNET -RELATED SOCIAL ISSUES - HSI 2005 3597: 238-247. LECTURE NOTES IN COMPUTER SCIENCE
Ko, Y; Park, J; Seo, J. 2004. Improving text categorization using the importance of sentences. INFORMATION PROCESSING & MANAGEMENT 40 (1): 65-79.
Ko, Y; Seo, J. 2004. Using the feature projection technique based on a normalized voting method for text classification. INFORMATION PROCESSING & MANAGEMENT 40 (2): 191-208.
Kobayashi, M; Aono, M; Takeuchi, H; Samukawa, H. 2002. Matrix computations for information retrieval and major and outlier cluster detection. JOURNAL OF COMPUTATIONAL AND APPLIED MATHEMATICS 149 (1): 119-129.
Kobayashi, N; Inui, K; Matsumoto, Y; Tateishi, K; Fukushima, T. 2005. Collecting evaluative expressions for opinion extraction. NATURAL LANGUAGE PROCESSING - IJCNLP 2004 3248: 596-605. LECTURE NOTES IN COMPUTER SCIENCE
Kogan, J; Nicholas, C; Volkovich, V. 2003. Text mining with information - Theoretic clustering. COMPUTING IN SCIENCE & ENGINEERING 5 (6): 52-59.
Kolcz, A; Chowdhury, A. 2005. Improved naive Bayes for extremely skewed misclassification costs. KNOWLEDGE DISCOVERY IN DATABASES: PKDD 2005 3721: 561-568. LECTURE NOTES IN ARTIFICIAL INTELLIGENCE
Kolda, TG; O'Leary, DP. 1998. A semidiscrete matrix decomposition for latent semantic indexing in information retrieval. ACM TRANSACTIONS ON INFORMATION SYSTEMS 16 (4): 322-346.
Kolda, TG; O'Leary, DP. 2000. Algorithm 805: Computation and uses of the semidiscrete matrix decomposition. ACM TRANSACTIONS ON MATHEMATICAL SOFTWARE 26 (3): 415-435.
Kongovi, M; Guzman, JC; Dasigi, V. 2002. Text categorization: An experiment using phrases. ADVANCES IN INFORMATION REFTRIEVAL 2291: 213-228. LECTURE NOTES IN COMPUTER SCIENCE
Kontos, J; Malagardi, I; Alexandris, C; Bouligaraki, M. 2000. Greek verb semantic processing for stock market text mining. NATURAL LANGUAGE PROCESSING-NLP 2000, PROCEEDINGS 1835: 395-405. LECTURE NOTES IN ARTIFICIAL INTELLIGENCE

Kontostathis, A; Pottenger, WM. 2006. A framework for understanding Latent Semantic Indexing (LSI) performance. INFORMATION PROCESSING & MANAGEMENT 42 (1): 56-73.
Koster, CHA; Seutter, M. 2003. Taming wild phrases. ADVANCES IN INFORMATION RETRIEVAL 2633: 161-176. LECTURE NOTES IN COMPUTER SCIENCE
Koster, CHA; Seutter, M; Beney, J. 2003. Multi-classification of patent applications with Winnow. PERSPECTIVES OF SYSTEM INFORMATICS 2890: 546-555. LECTURE NOTES IN COMPUTER SCIENCE
Kostoff, R. 2001. The extraction of useful information from the biomedical literature. ACADEMIC MEDICINE 76 (12): 1265-1270.
Kostoff, RN. 1994. Research Impact Quantification. R & D MANAGEMENT 24 (3): 206-218.
Kostoff, RN. 1995. Research requirements for research impact assessment. RESEARCH POLICY 24 (6): 869-882.
Kostoff, RN. 1997. Citation analysis cross-field normalization: A new paradigm. SCIENTOMETRICS 39 (3): 225-230.
Kostoff, RN. 1998. The use and misuse of citation analysis in research evaluation - Comments on theories of citation?. SCIENTOMETRICS 43 (1): 27-43.
Kostoff, RN. 1999. Science and technology innovation. TECHNOVATION 19 (10): 593-604.
Kostoff, RN. 2002. Citation analysis of research performer quality. SCIENTOMETRICS 53 (1): 49-71.
Kostoff, RN. 2002. Overcoming specialization. BIOSCIENCE 52 (10): 937-941.
Kostoff, RN. 2003. Bilateral asymmetry prediction. MEDICAL HYPOTHESES 61 (2): 265-266.
Kostoff, RN. 2003. Role of technical literature in science and technology development and exploitation. JOURNAL OF INFORMATION SCIENCE 29 (3): 223-228.
Kostoff, RN; Bedford, CD; del Rio, JA; Cortes, HD; Karypis, G. 2004. Macromolecule mass spectrometry: Citation mining of user documents. JOURNAL OF THE AMERICAN SOCIETY FOR MASS SPECTROMETRY 15 (3): 281-287.
Kostoff, RN; Block, JA. 2005. Factor matrix text filtering and clustering. JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE AND TECHNOLOGY 56 (9): 946-968.

Kostoff, RN; Boylan, R; Simons, GR. 2004. Disruptive technology roadmaps. TECHNOLOGICAL FORECASTING AND SOCIAL CHANGE 71 (1-2): 141-159.
Kostoff, RN; Braun, T; Schubert, A; Toothman, DR; Humenik, JA. 2000. Fullerene data mining using bibliometrics and database tomography. JOURNAL OF CHEMICAL INFORMATION AND COMPUTER SCIENCES 40 (1): 19-39.
Kostoff, RN; del Rio, JA. 2001. The impact of physics research. PHYSICS WORLD 14 (6): 47-51.
Kostoff, RN; del Rio, JA; Cortes, HD; Smith, C; Smith, A; Wagner, C; Leydesdorff, L; Karypis, G; Malpohl, G; Tshiteya, R. 2005. The structure and infrastructure of Mexico's science and technology. TECHNOLOGICAL FORECASTING AND SOCIAL CHANGE 72 (7): 798-814.
Kostoff, RN; del Rio, JA; Humenik, JA; Garcia, EO; Ramirez, AM. 2001. Citation mining: Integrating text mining and bibliometrics for research user profiling. JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE AND TECHNOLOGY 52 (13): 1148-1156.
Kostoff, RN; DeMarco, RA. 2001. Extracting information from the literature by text mining. ANALYTICAL CHEMISTRY 73 (13): 370A-378A.
Kostoff, RN; Eberhart, HJ; Toothman, DR. 1997. Database tomography for information retrieval. JOURNAL OF INFORMATION SCIENCE 23 (4): 301-311.
Kostoff, RN; Eberhart, HJ; Toothman, DR. 1998. Database tomography for technical intelligence: A roadmap of the near-earth space science and technology literature. INFORMATION PROCESSING & MANAGEMENT 34 (1): 69-85.
Kostoff, RN; Eberhart, HJ; Toothman, DR. 1999. Hypersonic and supersonic flow roadmaps using bibliometrics and database tomography. JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE 50 (5): 427-447.
Kostoff, RN; Eberhart, HJ; Toothman, DR; Pellenbarg, R. 1997. Database Tomography for technical intelligence: Comparative roadmaps of the research impact assessment literature and the journal of the American Chemical Society. SCIENTOMETRICS 40 (1): 103-138.
Kostoff, RN; Geisler, E. 1999. Strategic management and implementation of textual data mining in government organizations. TECHNOLOGY ANALYSIS & STRATEGIC MANAGEMENT 11 (4): 493-525.
Kostoff, RN; Green, KA; Toothman, DR; Humenik, JA. 2000. Database tomography applied to an aircraft science and technology investment strategy. JOURNAL OF AIRCRAFT 37 (4): 727-730.

Kostoff, RN; Karpouzian, G; Malpohl, G. 2005. Text mining the global abrupt-wing-stall literature. JOURNAL OF AIRCRAFT 42 (3): 661-664.
Kostoff, RN; Martinez, WL. 2005. Is citation normalization realistic?. JOURNAL OF INFORMATION SCIENCE 31 (1): 57-61.
Kostoff, RN; Scaller, RR. 2001. Science and technology roadmaps. IEEE TRANSACTIONS ON ENGINEERING MANAGEMENT 48 (2): 132-143.
Kostoff, RN; Shlesinger, MF. 2005. CAB: Citation-assisted background. SCIENTOMETRICS 62 (2): 199-212.
Kostoff, RN; Shlesinger, MF; Malpohl, G. 2004. Fractals text mining using bibliometrics and database tomography. FRACTALS-COMPLEX GEOMETRY PATTERNS AND SCALING IN NATURE AND SOCIETY 12 (1): 1-16.
Kostoff, RN; Tshiteya, R; Pfeil, KM; Humenik, JA. 2002. Electrochemical power text mining using bibliometrics and database tomography. JOURNAL OF POWER SOURCES 110 (1): 163-176.
Kostoff, RN; Tshiteya, R; Pfeil, KM; Humenik, JA; Karypis, G. 2005. Power source roadmaps using bibliometrics and database tomography. ENERGY 30 (5): 709-730.
Kotis, K; Vouros, GA; Stergiou, K. 2004. Capturing semantics towards automatic coordination of domain ontologies. ARTIFICIAL INTELLIGENCE: METHODOLOGY, SYSTEMS, AND APPLICATIONS, PROCEEDINGS 3192: 22-32. LECTURE NOTES IN COMPUTER SCIENCE
Kotropoulos, C; Pitas, I. 2003. Segmentation of ultrasonic images using Support Vector Machines. PATTERN RECOGNITION LETTERS 24 (4-5): 715-727.
Kou, HZ; Napoli, A; Toussaint, Y. 2005. Application of text categorization to astronomy field. NATURAL LANGUAGE PROCESSING AND INFORMATION SYSTEMS, PROCEEDINGS 3513: 32-43. LECTURE NOTES IN COMPUTER SCIENCE
KOULOPOULOS, TM. 1992. DOCUMENT CLUSTERING. BYTE 17 (6): 272-273.
Kowalczyk, A; Raskutti, B; Ferra, H. 2004. Exploring potential of leave-one-out estimator for calibration of SVM in text mining. ADVANCES IN KNOWLEDGE DISCOVERY AND DATA MINING, PROCEEDINGS 3056: 361-372. LECTURE NOTES IN ARTIFICIAL INTELLIGENCE
Krishnapuram, R; Joshi, A; Nasraoui, O; Yi, LY. 2001. Low-complexity fuzzy relational clustering algorithms for Web mining. IEEE TRANSACTIONS ON FUZZY SYSTEMS 9 (4): 595-607.

Krogel, MA; Scheffer, T. 2004. Multi-relational learning, text mining, and semi-supervised learning for functional genomics. MACHINE LEARNING 57 (1-2): 61-81.
Kruengkrai, C; Sornlertlamvanich, V; Isahara, H. 2005. Document clustering using linear partitioning hyperplanes and reallocation. INFORMATION RETRIEVAL TECHNOLOGY 3411: 36-47. LECTURE NOTES IN COMPUTER SCIENCE
Kruger, O; Ladewig, J; Koster, K; Ragg, H. 2002. Widespread occurrence of serpin genes with multiple reactive centre-containing exon cassettes in insects and nematodes. GENE 293 (1-2): 97-105.
Ku, S; Lee, B; Lee, D. 2002. TWIMC: An anonymous recipient E-mail system. DEVELOPMENTS IN APPLIED ARTIFICAIL INTELLIGENCE, PROCEEDINGS 2358: 363-372. LECTURE NOTES IN ARTIFICIAL INTELLIGENCE
Kuflik, T; Boger, Z; Shoval, P. 2006. Filtering search results using an optimal set of terms identified by an artificial neural network. INFORMATION PROCESSING & MANAGEMENT 42 (2): 469-483.
Kugo, A; Yoshikawa, H; Shimoda, H; Wakabayashi, Y. 2005. Text mining analysis of public comments regarding high-level radioactive waste disposal. JOURNAL OF NUCLEAR SCIENCE AND TECHNOLOGY 42 (9): 755-767.
Kuhnhold, M. 2000. The concept of "text mining". WIRTSCHAFTSINFORMATIK 42 (2): 175-179.
Kuo, HKJ; Lee, CH. 2003. Discriminative training of natural language call routers. IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING 11 (1): 24-35.
Kurimo, M. 2002. Thematic indexing of spoken documents by using self-organizing maps. SPEECH COMMUNICATION 38 (1-2): 29-45.
Kusumura, Y; Hijikata, Y; Nishida, S. 2004. NTM-Agent: Text mining agent for net auction. IEICE TRANSACTIONS ON INFORMATION AND SYSTEMS E87D (6): 1386-1396.
Kwon, OW; Lee, JH. 2003. Text categorization based on k-nearest neighbor approach for Web site classification. INFORMATION PROCESSING & MANAGEMENT 39 (1): 25-44.
Lagus, K. 2002. Text retrieval using self-organized document maps. NEURAL PROCESSING LETTERS 15 (1): 21-29.
Lagus, K; Honkela, T; Kaski, S; Kohonen, T. 1999. Websom for textual data mining. ARTIFICIAL INTELLIGENCE REVIEW 13 (5-6): 345-364.

Lagus, K; Kaski, S; Kohonen, T. 2004. Mining massive document collections by the WEBSOM method. INFORMATION SCIENCES 163 (1-3): 135-156.
Lakshminarayan, C; Yu, QF; Benson, A. 2005. Improving customer experience via text mining. DATABASES IN NETWORKED INFORMATION SYSTEMS, PROCEEDINGS 3433: 288-299. LECTURE NOTES IN COMPUTER SCIENCE
Lam, W; Han, YQ. 2003. Automatic textual document categorization based on generalized instance sets and a metamodel. IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE 25 (5): 628-633.
Lam, W; Ruiz, M; Srinivasan, P. 1999. Automatic text categorization and its application to text retrieval. IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING 11 (6): 865-879.
Landau, D; Feldman, R; Aumann, Y; Fresko, M; Lindell, Y; Lipshtat, O; Zamir, O. 1998. TextVis: An integrated visual environment for text mining. PRINCIPLES OF DATA MINING AND KNOWLEDGE DISCOVERY 1510: 56-64. LECTURE NOTES IN ARTIFICIAL INTELLIGENCE
Lanquillon, C. 2000. Partially supervised text classification: Combining labeled and unlabeled documents using an EM-like scheme. MACHINE LEARNING: ECML 2000 1810: 229-237. LECTURE NOTES IN ARTIFICIAL INTELLIGENCE
Lapalut, S. 1996. Text clustering to help knowledge acquisition from documents. ADVANCES IN KNOWLEDGE ACQUISITION 1076: 115-130. LECTURE NOTES IN ARTIFICIAL INTELLIGENCE
Lappe, M; Holm, L. 2005. Algorithms for protein interaction networks. BIOCHEMICAL SOCIETY TRANSACTIONS 33: 530-534, Part 3.
Lau, RYK. 2003. Belief revision for adaptive recommender agents in E-commerce. INTELLIGENT DATA ENGINEERING AND AUTOMATED LEARNING 2690: 99-103. LECTURE NOTES IN COMPUTER SCIENCE
Lau, RYK; van den Brand, P. 2003. Belief revision and text mining for adaptive recommender agents. FOUNDATIONS OF INTELLIGENT SYSTEMS 2871: 226-230. LECTURE NOTES IN ARTIFICIAL INTELLIGENCE
Lebart, L. 1998. Text mining in different languages. APPLIED STOCHASTIC MODELS AND DATA ANALYSIS 14 (4): 323-334.
Lee, B; Park, Y. 2003. An e-mail monitoring system for detecting outflow of confidential documents. INTELLIGENCE AND SECURITY

INFORMATICS, PROCEEDINGS 2665: 371-374. LECTURE NOTES IN COMPUTER SCIENCE
Lee, CH; Yang, HC. 2003. A multilingual text mining approach based on self-organizing maps. APPLIED INTELLIGENCE 18 (3): 295-310.
Lee, CK; Lee, GG. 2006. Information gain and divergence-based feature selection for machine learning-based text categorization. INFORMATION PROCESSING & MANAGEMENT 42 (1): 155-165.
Lee, KC; Kang, SS; Hahn, KS. 2005. A term weighting approach for text categorization. INFORMATION RETRIEVAL TECHNOLOGY, PROCEEDINGS 3689: 673-678. LECTURE NOTES IN COMPUTER SCIENCE
Lee, KH; Kay, J; Kang, BH. 2003. Active learning: Applying RinSCut thresholding strategy to uncertainty sampling. AI 2003: ADVANCES IN ARTIFICIAL INTELLIGENCE 2903: 922-932. LECTURE NOTES IN ARTIFICIAL INTELLIGENCE
Lee, KS; Kageura, K. 2003. Incorporating virtual relevant documents for learning in text categorization. DIGITAL LIBRARIES: TECHNOLOGY AND MANAGEMENT OF INDIGENOUS KNOWLEDGE FOR GLOBAL ACCESS 2911: 62-72. LECTURE NOTES IN COMPUTER SCIENCE
Lee, KS; Kageura, K; Choi, KS. 2004. Implicit ambiguity resolution using incremental clustering in cross-language information retrieval. INFORMATION PROCESSING & MANAGEMENT 40 (1): 145-159.
Lee, KS; Park, YC; Choi, KS. 2001. Re-ranking model based on document clusters. INFORMATION PROCESSING & MANAGEMENT 37 (1): 1-14.
Lee, MD; Corlett, EY. 2003. Sequential sampling models of human text classification. COGNITIVE SCIENCE 27 (2): 159-193.
LEHNERT, W; SODERLAND, S; ARONOW, D; FENG, FF; SHMUELI, A. 1995. INDUCTIVE TEXT CLASSIFICATION FOR MEDICAL APPLICATIONS. JOURNAL OF EXPERIMENTAL & THEORETICAL ARTIFICIAL INTELLIGENCE 7 (1): 49-80.
Leopold, E; Kindermann, J. 2002. Text categorization with support vector machines. How to represent texts in input space ?. MACHINE LEARNING 46 (1-3): 423-444.
Lertnattee, V; Theeramunkong, T. 2003. Term-length normalization for centroid-based text categorization. KNOWLEDGE-BASED INTELLIGENT INFORMATION AND ENGINEERING SYSTEMS, PT 1, PROCEEDINGS 2773: 850-856. LECTURE NOTES IN ARTIFICIAL INTELLIGENCE

Lertnattee, V; Theeramunkong, T. 2004. Effect of term distributions on centroid-based text categorization. INFORMATION SCIENCES 158: 89-115.
Lertnattee, V; Theeramunkong, T. 2004. Multidimensional text classification for drug information. IEEE TRANSACTIONS ON INFORMATION TECHNOLOGY IN BIOMEDICINE 8 (3): 306-312.
Lertnattee, V; Theeramunkong, T. 2004. Parallel text categorization for multi-dimensional data. PARALLEL AND DISTRIBUTED COMPUTING: APPLICATIONS AND TECHNOLOGIES, PROCEEDINGS 3320: 38-41. LECTURE NOTES IN COMPUTER SCIENCE
Leslie, C; Kuang, R. 2003. Fast kernels for inexact string matching. LEARNING THEORY AND KERNEL MACHINES 2777: 114-128. LECTURE NOTES IN ARTIFICIAL INTELLIGENCE
Letsche, TA; Berry, MW. 1997. Large-scale information retrieval with latent semantic indexing. INFORMATION SCIENCES 100 (1-4): 105-137.
Leuski, A; Allan, J. 2004. Interactive information retrieval using clustering and spatial proximity. USER MODELING AND USER-ADAPTED INTERACTION 14 (2-3): 259-288.
Lewis, KN; Robinson, MD; Hughes, TR; Hogue, CWV. 2004. MyMED: A database system for biomedical research on MEDLINE data. IBM SYSTEMS JOURNAL 43 (4): 756-767.
Li, BL; Chen, YZ; Bai, XJ; Yu, SW. 2003. Experimental study on representing units in Chinese text categorization. COMPUTATIONAL LINGUISTICS AND INTELLIGENT TEXT PROCESSING, PROCEEDINGS 2588: 602-614. LECTURE NOTES IN COMPUTER SCIENCE
Li, H; Li, C. 2004. Word translation disambiguation using bilingual bootstrapping. COMPUTATIONAL LINGUISTICS 30 (1): 1-22.
Li, H; Yamanishi, K. 2002. Text classification using ESC-based stochastic decision lists. INFORMATION PROCESSING & MANAGEMENT 38 (3): 343-361.
Li, H; Yamanishi, K. 2003. Topic analysis using a finite mixture model. INFORMATION PROCESSING & MANAGEMENT 39 (4): 521-541.
Li, HP; Doermann, D; Kia, O. 1999. Text extraction, enhancement and OCR in digital video. DOCUMENT ANALYSIS SYSTEMS: THEORY AND PRACTICE 1655: 363-377. LECTURE NOTES IN COMPUTER SCIENCE
Li, RL; Tao, XP; Tang, L; Hu, YF. 2004. Using maximum entropy model for Chinese text categorization. ADVANCED WEB TECHNOLOGIES

AND APPLICATIONS 3007: 578-587. LECTURE NOTES IN COMPUTER SCIENCE
Li, XG; Yu, G; Wang, DL. 2005. MMPClust: A skew prevention algorithm for model-based document clustering. DATABASE SYSTEMS FOR ADVANCED APPLICATIONS, PROCEEDINGS 3453: 536-547. LECTURE NOTES IN COMPUTER SCIENCE
Li, XG; Yu, G; Wang, DL; Bao, YB. 2005. ESPClust: An effective skew prevention method for model-based document clustering. COMPUTATIONAL LINGUISTICS AND INTELLIGENT TEXT PROCESSING 3406: 735-745. LECTURE NOTES IN COMPUTER SCIENCE
Li, XL; Shi, ZZ. 2002. Innovating web page classification through reducing noise. JOURNAL OF COMPUTER SCIENCE AND TECHNOLOGY 17 (1): 9-17.
Liang, BY; Tang, J; Li, JZ; Wang, KH. 2004. Keyword extraction based peer clustering. GRID AND COOPERATIVE COMPUTING GCC 2004, PROCEEDINGS 3251: 827-830. LECTURE NOTES IN COMPUTER SCIENCE
Liao, SS; Jiang, MH. 2005. An improved method of feature selection based on concept attributes in text classification. ADVANCES IN NATURAL COMPUTATION, PT 1, PROCEEDINGS 3610: 1140-1149. LECTURE NOTES IN COMPUTER SCIENCE
Liao, YH; Vemuri, VR. 2002. Use of K-Nearest Neighbor classifier for intrusion detection. COMPUTERS & SECURITY 21 (5): 439-448.
LIDDY, ED; PAIK, WJ; YU, ES. 1994. TEXT CATEGORIZATION FOR MULTIPLE USERS BASED ON SEMANTIC FEATURES FROM A MACHINE-READABLE DICTIONARY. ACM TRANSACTIONS ON INFORMATION SYSTEMS 12 (3): 278-295.
Lim, HS. 2004. Improving kNN based text classification with well estimated parameters. NEURAL INFORMATION PROCESSING 3316: 516-523. LECTURE NOTES IN COMPUTER SCIENCE
Lin, CT; Chen, HC; Nunamaker, JF. 1999. Verifying the proximity and size hypothesis for self-organizing maps. JOURNAL OF MANAGEMENT INFORMATION SYSTEMS 16 (3): 57-70.
Lin, HF; Zhan, XG; Yao, TS. 1999. Example-based Chinese text filtering model. INTERNET APPLICATIONS 1749: 415-420. LECTURE NOTES IN COMPUTER SCIENCE
Lin, TY; Chiang, IJ. 2005. A simplicial complex, a hypergraph, structure in the latent semantic space of document clustering. INTERNATIONAL JOURNAL OF APPROXIMATE REASONING 40 (1-2): 55-80.

Lindsay, RK; Gordon, MD. 1999. Literature-based discovery by lexical statistics. JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE 50 (7): 574-587.
Lindsey, CS; Stromberg, M. 2000. Image classification using the frequencies of simple features. PATTERN RECOGNITION LETTERS 21 (3): 265-268.
Liu, CL; Chang, CT; Ho, JH. 2004. Case instance generation and refinement for case-based criminal summary judgments in Chinese. JOURNAL OF INFORMATION SCIENCE AND ENGINEERING 20 (4): 783-800.
Liu, F; Jenssen, TK; Nygaard, V; Sack, J; Hovig, E. 2004. FigSearch: a figure legend indexing and classification system. BIOINFORMATICS 20 (16): 2880-2882.
Liu, F; Ma, FY; Li, ML; Huang, LP. 2004. A framework for semantic grid service discovery. WEB INFORMATION SYSTEMS - WISE 2004 WORKSHOPS, PROCEEDINGS 3307: 3-10. LECTURE NOTES IN COMPUTER SCIENCE
Liu, F; Ma, FY; Li, ML; Huang, LP. 2004. Distributed information retrieval based on hierarchical semantic overlay network. GRID AND COOPERATIVE COMPUTING GCC 2004, PROCEEDINGS 3251: 657-664. LECTURE NOTES IN COMPUTER SCIENCE
Liu, F; Ma, FY; Ye, YM; Li, ML; Yu, JD. 2005. IglooG: A distributed web crawler based on grid service. WEB TECHNOLOGIES RESEARCH AND DEVELOPMENT - APWEB 2005 3399: 207-216. LECTURE NOTES IN COMPUTER SCIENCE
Liu, F; Zhang, WJ; Ma, FY; Li, ML. 2004. SPSS: A case of semantic peer-to-peer search system. WEB INFORMATION SYSTEMS - WISE 2004, PROCEEDINGS 3306: 718-723. LECTURE NOTES IN COMPUTER SCIENCE
Liu, F; Zhang, WJ; Yu, S; Ma, FY; Li, ML. 2004. A peer-to-peer hypertext categorization using Directed Acyclic Graph Support Vector Machines. PARALLEL AND DISTRIBUTED COMPUTING: APPLICATIONS AND TECHNOLOGIES, PROCEEDINGS 3320: 54-57. LECTURE NOTES IN COMPUTER SCIENCE
Liu, FF; Jin, QL; Zhao, J; Xu, B. 2005. Bilingual chunk alignment based on interactional matching and probabilistic latent semantic indexing. NATURAL LANGUAGE PROCESSING - IJCNLP 2004 3248: 416-425. LECTURE NOTES IN COMPUTER SCIENCE
Liu, H; Huang, ST. 2003. A genetic semi-supervised fuzzy clustering approach to text classification. ADVANCES IN WEB-AGE INFORMATION MANAGEMENT, PROCEEDINGS 2762: 173-180. LECTURE NOTES IN COMPUTER SCIENCE

Liu, H; Huang, ST. 2003. Evolutionary semi-supervised fuzzy clustering. PATTERN RECOGNITION LETTERS 24 (16): 3105-3113.
Liu, JJ; Cutler, G; Li, WX; Pan, Z; Peng, SH; Hoey, T; Chen, LB; Ling, XFB. 2005. Multiclass cancer classification and biomarker discovery using GA-based algorithms. BIOINFORMATICS 21 (11): 2691-2697.
Liu, RL; Lin, WJ. 2003. Mining for interactive identification of users' information needs. INFORMATION SYSTEMS 28 (7): 815-833.
Liu, RL; Lin, WJ. 2005. Incremental mining of information interest for personalized web scanning. INFORMATION SYSTEMS 30 (8): 630-648.
Liu, T; Guo, J. 2005. Text similarity computing based on standard deviation. ADVANCES IN INTELLIGENT COMPUTING, PT 1, PROCEEDINGS 3644: 456-464. LECTURE NOTES IN COMPUTER SCIENCE
Liu, XM; Yin, JW; Dong, JX; Ghafoor, MA. 2005. An improved FloatBoost algorithm for Naive Bayes text classification. ADVANCES IN WEB-AGE INFORMATION MANAGEMENT, PROCEEDINGS 3739: 162-171. LECTURE NOTES IN COMPUTER SCIENCE
Liu, XW; He, PL. 2005. A study on text clustering algorithms based on frequent term sets. ADVANCED DATA MINING AND APPLICATIONS, PROCEEDINGS 3584: 347-354. LECTURE NOTES IN ARTIFICIAL INTELLIGENCE
Liu, XZ; Chen, M; Yang, GW. 2004. Latent semantic indexing in peer-to-peer networks. ORGANIC AND PERVASIVE COMPUTING - ARCS 2004 2981: 63-77. LECTURE NOTES IN COMPUTER SCIENCE
Liu, Y; Carbonell, J; Jin, R. 2003. A new pairwise ensemble approach for text classification. MACHINE LEARNING: ECML 2003 2837: 277-288. LECTURE NOTES IN ARTIFICIAL INTELLIGENCE
Liu, Y; Loh, HT; Tor, SB. 2005. Comparison of extreme learning machine with support vector machine for text classification. INNOVATIONS IN APPLIED ARTIFICIAL INTELLIGENCE 3533: 390-399. LECTURE NOTES IN ARTIFICIAL INTELLIGENCE
LOCHBAUM, KE; STREETER, LA. 1989. COMPARING AND COMBINING THE EFFECTIVENESS OF LATENT SEMANTIC INDEXING AND THE ORDINARY VECTOR-SPACE MODEL FOR INFORMATION-RETRIEVAL. INFORMATION PROCESSING & MANAGEMENT 25 (6): 665-676.
Lodhi, H; Saunders, C; Shawe-Taylor, J; Cristianini, N; Watkins, C. 2002. Text classification using string kernels. JOURNAL OF MACHINE LEARNING RESEARCH 2 (3): 419-444.

Loh, S; De Oliveira, JPM; Gameiro, MA. 2003. Knowledge discovery in texts for constructing decision support systems. APPLIED INTELLIGENCE 18 (3): 357-366.
Lorensuhewa, A; Pham, B; Geva, S. 2003. Style recognition using keyword analysis. MINING MULTIMEDIA AND COMPLEX DATA 2797: 266-280. LECTURE NOTES IN ARTIFICIAL INTELLIGENCE
Losee, RM; Church, L. 2005. Are two document clusters better than one? The cluster performance question for information retrieval. JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE AND TECHNOLOGY 56 (1): 106-108.
Losiewicz, P; Oard, DW; Kostoff, RN. 2000. Textual data mining to support science and technology management. JOURNAL OF INTELLIGENT INFORMATION SYSTEMS 15 (2): 99-119.
Lu, HJ; Luo, Q; Shun, YK. 2003. Extending a Web Browser with client-side mining. WEB TECHNOLOGIES AND APPLICATIONS 2642: 166-177. LECTURE NOTES IN COMPUTER SCIENCE
Lu, JJ; Xu, BW; Jiang, JX. 2004. Generating different semantic spaces for document classification. CONTENT COMPUTING, PROCEEDINGS 3309: 430-436. LECTURE NOTES IN COMPUTER SCIENCE
Lu, JJ; Xu, BW; Jiang, JX; Kang, DZ. 2004. Non-negative matrix factorization for filtering Chinese document. COMPUTATIONAL SCIENCE - ICCS 2004, PT 2, PROCEEDINGS 3037: 113-120. LECTURE NOTES IN COMPUTER SCIENCE
Lu, W; Kan, MY. 2005. Supervised categorization of JavaScript (TM) using program analysis features. INFORMATION RETRIEVAL TECHNOLOGY, PROCEEDINGS 3689: 160-173. LECTURE NOTES IN COMPUTER SCIENCE
Lu, WH; Chien, LF; Lee, HJ. 2004. Anchor text mining for translation of Web queries: A transitive translation approach. ACM TRANSACTIONS ON INFORMATION SYSTEMS 22 (2): 242-269.
Luehrs, R; Pavon, J; Schneider, M. 2003. DEMOS tools for online discussion and decision making. WEB ENGINEERING, PROCEEDINGS 2722: 525-528. LECTURE NOTES IN COMPUTER SCIENCE
Luo, DS; Wang, XH; Wu, XH; Chi, HS. 2005. Learning outliers to refine a corpus for Chinese webpage categorization. ADVANCES IN NATURAL COMPUTATION, PT 1, PROCEEDINGS 3610: 167-178. LECTURE NOTES IN COMPUTER SCIENCE
Luo, X; Zincir-Heywood, AN. 2004. Analyzing the temporal sequences for text categorization. KNOWLEDGE-BASED INTELLIGENT INFORMATION AND ENGINEERING SYSTEMS, PT 3,

PROCEEDINGS 3215: 498-505. LECTURE NOTES IN ARTIFICIAL INTELLIGENCE
Luo, X; Zincir-Heywood, AN. 2005. Evaluation of two systems on multi-class multi-label document classification. FOUNDATIONS OF INTELLIGENT SYSTEMS, PROCEEDINGS 3488: 161-169. LECTURE NOTES IN COMPUTER SCIENCE
Lussner, W. 2000. Knowledge management technologies. NFD INFORMATION-WISSENSCHAFT UND PRAXIS 51 (6): 364-366.
Maarek, YS; BenShaul, IZ. 1996. Automatically organizing bookmarks per contents. COMPUTER NETWORKS AND ISDN SYSTEMS 28 (7-11): 1321-1333.
Mack, R; Mukherjea, S; Soffer, A; Uramoto, N; Brown, E; Coden, A; Cooper, J; Inokuchi, A; Iyer, B; Mass, Y; Matsuzawa, H; Subramaniam, LV. 2004. Text analytics for life science using the unstructured information management architecture. IBM SYSTEMS JOURNAL 43 (3): 490-515.
MACLEOD, KJ; ROBERTSON, W. 1991. A NEURAL ALGORITHM FOR DOCUMENT CLUSTERING. INFORMATION PROCESSING & MANAGEMENT 27 (4): 337-346.
MacMullen, WJ; Vaughan, KTL; Moore, ME. 2004. Planning bioinformatics education and information services in an academic health sciences library. COLLEGE & RESEARCH LIBRARIES 65 (4): 320-333.
Macskassy, SA; Hirsh, H; Banerjee, A; Dayanik, AA. 2003. Converting numerical classification into text classification. ARTIFICIAL INTELLIGENCE 143 (1): 51-77.
Maderlechner, G; Suda, P; Bruckner, T. 1997. Classification of documents by form and content. PATTERN RECOGNITION LETTERS 18 (11-13): 1225-1231.
Makagonov, P; Sboychakov, K. 2001. Software for creating domain-oriented dictionaries and document clustering in full-text databases. COMPUTATIONAL LINGUISTICS AND INTELLIGENT TEXT PROCESSING 2004: 454-456. LECTURE NOTES IN COMPUTER SCIENCE
Makrehchi, M; Kamel, MS. 2005. Text classification using small number of features. MACHINE LEARNING AND DATA MINING IN PATTERN RECOGNITION, PROCEEDINGS 3587: 580-589. LECTURE NOTES IN ARTIFICIAL INTELLIGENCE
Manjula, D; Geetha, TV. 2004. Description logic based knowledge representation for information extraction and query processing. IETE JOURNAL OF RESEARCH 50 (6): 437-441.

Marcus, A; Maletic, JI; Sergeyev, A. 2005. Recovery of traceability links between software documentation and source code. INTERNATIONAL JOURNAL OF SOFTWARE ENGINEERING AND KNOWLEDGE ENGINEERING 15 (5): 811-836.
Marques, NC; Braud, A. 2003. Mining generalized character n-grams in large corpora. PROGRESS IN ARTIFICIAL INTELLIGENCE 2902: 419-423. LECTURE NOTES IN ARTIFICIAL INTELLIGENCE
Martin, DMA; Hill, P; Barton, GJ; Flavell, AJ. 2003. Visual representation of database search results: the RHIMS Plot. BIOINFORMATICS 19 (8): 1037-1038.
Martin, EPG; Bremer, EG; Guerin, MC; DeSesa, C; Jouve, O. 2004. Analysis of protein/protein interactions through biomedical literature: Text mining of abstracts vs. text mining of full text articles. KNOWLEDGE EXPLORATION IN LIFE SCIENCE INFORMATICS, PROCEEDINGS 3303: 96-108. LECTURE NOTES IN ARTIFICIAL INTELLIGENCE
Martinez-Trinidad, JF; Beltran-Martinez, B; Ruiz-Shulcloper, J. 2000. A tool to discover the main themes in a Spanish or English document. EXPERT SYSTEMS WITH APPLICATIONS 19 (4): 319-327.
Martin-Merino, M; Munoz, A. 2001. Self Organizing Map and Sammon Mapping for asymmetric proximities. ARTIFICIAL NEURAL NETWORKS-ICANN 2001, PROCEEDINGS 2130: 429-435. LECTURE NOTES IN COMPUTER SCIENCE
Martin-Merino, M; Munoz, A. 2005. Extending the SOM algorithm to visualize word relationships. ADVANCES IN INTELLIGENT DATA ANALYSIS VI, PROCEEDINGS 3646: 228-238. LECTURE NOTES IN COMPUTER SCIENCE
Martin-Merino, M; Munoz, A. 2005. Visualizing asymmetric proximities with SOM and MDS models. NEUROCOMPUTING 63: 171-192.
Martin-Valdivia, MT; Garcia-Vega, M; Urena-Lopez, LA. 2003. LVQ for text categorization using a multilingual linguistic resource. NEUROCOMPUTING 55 (3-4): 665-679.
Marton, Y; Wu, N; Hellerstein, L. 2005. On compression-based text classification. ADVANCES IN INFORMATION RETRIEVAL 3408: 300-314. LECTURE NOTES IN COMPUTER SCIENCE
Massey, L. 2003. On the quality of ART1 text clustering. NEURAL NETWORKS 16 (5-6): 771-778.
Masuyama, T; Nakagawa, H. 2003. Cascaded feature selection in SVMs text categorization. COMPUTATIONAL LINGUISTICS AND INTELLIGENT TEXT PROCESSING, PROCEEDINGS 2588: 588-591. LECTURE NOTES IN COMPUTER SCIENCE

Masuyama, T; Nakagawa, H. 2004. Two step POS selection for SVM based text categorization. IEICE TRANSACTIONS ON INFORMATION AND SYSTEMS E87D (2): 373-379.
Mather, LA. 2000. A linear algebra measure of cluster quality. JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE 51 (7): 602-613.
Matsuda, Y; Yamaguchi, K. 2005. An efficient MDS algorithm for the analysis of massive document collections. KNOWLEDGE-BASED INTELLIGENT INFORMATION AND ENGINEERING SYSTEMS, PT 2, PROCEEDINGS 3682: 1015-1021. LECTURE NOTES IN ARTIFICIAL INTELLIGENCE
Matsumoto, S; Takamura, H; Okumura, M. 2005. Sentiment classification using word sub-sequences and dependency sub-trees. ADVANCES IN KNOWLEDGE DISCOVERY AND DATA MINING, PROCEEDINGS 3518: 301-311. LECTURE NOTES IN ARTIFICIAL INTELLIGENCE
Matsuzawa, H; Fukuda, T. 2000. Mining structured association patterns from databases. KNOWLEDGE DISCOVERY AND DATA MINING, PROCEEDINGS 1805: 233-244. LECTURE NOTES IN ARTIFICIAL INTELLIGENCE
Mavroeidis, D; Tsatsaronis, G; Vazirgiannis, M; Theobald, M; Weikum, G. 2005. Word sense disambiguation for exploiting hierarchical thesauri in text classification. KNOWLEDGE DISCOVERY IN DATABASES: PKDD 2005 3721: 181-192. LECTURE NOTES IN ARTIFICIAL INTELLIGENCE
McCallum, AK; Nigam, K; Rennie, J; Seymore, K. 2000. Automating the construction of internet portals with machine learning. INFORMATION RETRIEVAL 3 (2): 127-163.
McDonald, DM; Chen, HC; Su, H; Marshall, BB. 2004. Extracting gene pathway relations using a hybrid grammar: the Arizona Relation Parser. BIOINFORMATICS 20 (18): 3370-3378.
Meier, M; Beckh, M. 2000. Text mining. WIRTSCHAFTSINFORMATIK 42 (2): 165-167.
Menon, R; Tong, LH; Sathiyakeerthi, S; Brombacher, A. 2003. Automated text classification for fast feedback - Investigating the effects of document representation. KNOWLEDGE-BASED INTELLIGNET INFORMATION AND ENGINEERING SYSTEMS, PT 2, PROCEEDINGS 2774: 1008-1014. LECTURE NOTES IN ARTIFICIAL INTELLIGENCE
Menon, R; Tong, LH; Sathiyakeerthi, S; Brombacher, A; Leong, C. 2004. The needs and benefits of applying textual data mining within the product

development process. QUALITY AND RELIABILITY ENGINEERING INTERNATIONAL 20 (1): 1-15.
Merkl, D. 1997. Exploration of document collections with self-organizing maps: A novel approach to similarity representation. PRINCIPLES OF DATA MINING AND KNOWLEDGE DISCOVERY 1263: 101-111. LECTURE NOTES IN ARTIFICIAL INTELLIGENCE
Merkl, D. 1998. Text classification with self-organizing maps: Some lessons learned. NEUROCOMPUTING 21 (1-3): 61-77.
Metzger, J; Schillo, M; Fischer, K. 2003. A multiagent-based peer-to-peer network in Java for distributed spam filtering. MULTI-AGENT SYSTEMS AND APPLICATIONS III, PROCEEDINGS 2691: 616-625. LECTURE NOTES IN ARTIFICIAL INTELLIGENCE
Miller, DJ; Browning, J. 2003. A mixture model and EM-based algorithm for class discovery, robust classification, and outlier rejection in mixed labeled/unlabeled data sets. IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE 25 (11): 1468-1483.
Miller, MH. 1997. A method for representing search results in three dimensions. JOURNAL OF THE AMERICAN MEDICAL INFORMATICS ASSOCIATION: 533-537, Suppl. S.
Milward, D; Bjareland, M; Hayes, W; Maxwell, M; Oberg, L; Tilford, N; Thomas, J; Hale, R; Knight, S; Barnes, JE. 2005. Ontology-based interactive information extraction from scientific abstracts. COMPARATIVE AND FUNCTIONAL GENOMICS 6 (1-2): 67-71.
Missikoff, M; Velardi, P; Fabriani, P. 2003. Text mining techniques to automatically enrich a domain ontology. APPLIED INTELLIGENCE 18 (3): 323-340.
Miyamoto, S. 2003. Information clustering based on fuzzy multisets. INFORMATION PROCESSING & MANAGEMENT 39 (2): 195-213.
Miyamoto, S; Mizutani, K. 2004. Fuzzy multiset model and methods of nonlinear document clustering for information retrieval. MODELING DECISIONS FOR ARTIFICIAL INTELLIGENCE, PROCEEDINGS 3131: 273-283. LECTURE NOTES IN ARTIFICIAL INTELLIGENCE
Mladenic, D; Grobelnik, M. 2003. Feature selection on hierarchy of web documents. DECISION SUPPORT SYSTEMS 35 (1): 45-87.
Modha, DS; Spangler, WS. 2003. Feature weighting in k-means clustering. MACHINE LEARNING 52 (3): 217-237.
Moens, MF; Dumortier, J. 2000. Text categorization: the assignment of subject descriptors to magazine articles. INFORMATION PROCESSING & MANAGEMENT 36 (6): 841-861.

Mons, B. 2005. Which gene did you mean?. BMC BIOINFORMATICS 6: art. no.-142.
Montanes, E; Combarro, EF; Diaz, I; Ranilla, J. 2005. Towards automatic and optimal Filtering Levels for Feature Selection in Text Categorization. ADVANCES IN INTELLIGENT DATA ANALYSIS VI, PROCEEDINGS 3646: 239-248. LECTURE NOTES IN COMPUTER SCIENCE
Montanes, E; Combarro, EF; Diaz, I; Ranilla, J; Quevedo, JR. 2004. Words as rules: Feature selection in text categorization. COMPUTATIONAL SCIENCE - ICCS 2004, PT 1, PROCEEDINGS 3036: 666-669. LECTURE NOTES IN COMPUTER SCIENCE
Montanes, E; Diaz, I; Ranilla, J; Combarro, EF; Fernandez, J. 2005. Scoring and selecting terms for text categorization. IEEE INTELLIGENT SYSTEMS 20 (3): 40-47.
Montanes, E; Fernandez, J; Diaz, I; Combarro, EF; Ranilla, J. 2003. Measures of rule quality for feature selection in Text Categorization. ADVANCES IN INTELLIGENT DATA ANALYSIS V 2810: 589-598. LECTURE NOTES IN COMPUTER SCIENCE
Montanes, E; Quevedo, JR; Diaz, I. 2003. A wrapper approach with support vector machines for text categorization. COMPUTATIONAL METHODS IN NEURAL MODELING, PT 1 2686: 230-237. LECTURE NOTES IN COMPUTER SCIENCE
Montes-y-Gomez, M; Gelbukh, A; Lopez-Lopez, A. 2001. A statistical approach to the discovery of ephemeral associations among news topics. DATABASE AND EXPERT SYSTEMS APPLICATIONS 2113: 491-500. LECTURE NOTES IN COMPUTER SCIENCE
Montes-y-Gomez, M; Gelbukh, A; Lopez-Lopez, A. 2002. Text mining at detail level using conceptual graphs. CONCEPTUAL STRUCTURES: INTEGRATION AND INTERFACES, PROCEEDINGS 2393: 122-136. LECTURE NOTES IN ARTIFICIAL INTELLIGENCE
Montes-y-Gomez, M; Gelbukh, A; Lopez-Lopez, A; BaezaYates, R. 2001. Flexible comparison of conceptual graphs. DATABASE AND EXPERT SYSTEMS APPLICATIONS 2113: 102-111. LECTURE NOTES IN COMPUTER SCIENCE
Montes-y-Gomez, M; Perez-Coutino, M; Villasenor-Pineda, L; Lopez-Lopez, A. 2004. Contextual exploration of text collections. COMPUTATIONAL LINGUISTICS AND INTELLIGENT TEXT PROCESSING 2945: 488-497. LECTURE NOTES IN COMPUTER SCIENCE

Morgan, AA; Hirschman, L; Colosimo, M; Yeh, AS; Colombe, JB. 2004. Gene name identification and normalization using a model organism database. JOURNAL OF BIOMEDICAL INFORMATICS 37 (6): 396-410.
Morik, K; Kopcke, H. 2005. Features for learning local patterns in time-stamped data. LOCAL PATTERN DETECTION 3539: 98-114. LECTURE NOTES IN ARTIFICIAL INTELLIGENCE
Morita, K; Kadoya, Y; Atlam, ES; Fuketa, M; Kashiji, S; Aoe, J. 2004. A method of extracting and evaluating popularity and unpopularity for natural language expressions. KNOWLEDGE-BASED INTELLIGENT INFORMATION AND ENGINEERING SYSTEMS, PT 1, PROCEEDINGS 3213: 567-574. LECTURE NOTES IN COMPUTER SCIENCE
Morizet-Mahoudeaux, P; Bachimont, B. 2005. Indexing and mining audiovisual data. ACTIVE MINING 3430: 34-58. LECTURE NOTES IN ARTIFICIAL INTELLIGENCE
Morris, S; DeYong, C; Wu, Z; Salman, S; Yemenu, D. 2002. DIVA: a visualization system for exploring document databases for technology forecasting. COMPUTERS & INDUSTRIAL ENGINEERING 43 (4): 841-862.
Moschitti, A. 2003. A study on optimal parameter tuning for Rocchio text classifier. ADVANCES IN INFORMATION RETRIEVAL 2633: 420-435. LECTURE NOTES IN COMPUTER SCIENCE
Moschitti, A; Basili, R. 2004. Complex linguistic features for text classification: A comprehensive study. ADVANCES IN INFORMATION RETRIEVAL, PROCEEDINGS 2997: 181-196. LECTURE NOTES IN COMPUTER SCIENCE
Mostafa, J; Mukhopadhyay, S; Lam, W; Palakal, M. 1997. A multilevel approach to intelligent information filtering: Model, system, and evaluation. ACM TRANSACTIONS ON INFORMATION SYSTEMS 15 (4): 368-399.
Moyotl-Hernandez, E; Jimenez-Salazar, H. 2005. Enhancement of DTP feature selection method for text categorization. COMPUTATIONAL LINGUISTICS AND INTELLIGENT TEXT PROCESSING 3406: 719-722. LECTURE NOTES IN COMPUTER SCIENCE
Muller, HM; Kenny, EE; Sternberg, PW. 2004. Textpresso: An ontology-based information retrieval and extraction system for biological literature. PLOS BIOLOGY 2 (11): 1984-1998.
Muresan, G; Harper, DJ. 2004. Topic modeling for mediated access to very large document collections. JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE AND TECHNOLOGY 55 (10): 892-910.

Nakov, P. 2000. Web personalization using Extended Boolean operations with latent semantic indexing. ARTIFICIAL INTELLIGENCE: METHODOLOGY, SYSTEMS, APPLICATIONS, PROCEEDINGS 1904: 189-198. LECTURE NOTES IN ARTIFICIAL INTELLIGENCE
Nam, T; Lee, HG; Jeong, CY; Han, C. 2004. A harmful content protection in peer-to-peer networks. ARTIFICIAL INTELLIGENCE AND SIMULATION 3397: 617-626. LECTURE NOTES IN COMPUTER SCIENCE
Nanas, N; Uren, V; de Roeck, A; Domingue, J. 2004. Multi-topic information filtering with a single user profile. METHODS AND APPLICATIONS OF ARTIFICIAL INTELLIGENCE, PROCEEDINGS 3025: 400-409. LECTURE NOTES IN COMPUTER SCIENCE
Narayanasamy, V; Mukhopadhyay, S; Palakal, M; Potter, DA. 2004. TransMiner: Mining transitive associations among biological objects from text. JOURNAL OF BIOMEDICAL SCIENCE 11 (6): 864-873.
Narayanaswamy, M; Ravikumar, KE; Vijay-Shanker, K. 2005. Beyond the clause: extraction of phosphorylation information from medline abstracts. BIOINFORMATICS 21: I319-I327, Suppl. 1.
Nardiello, P; Sebastiani, F; Sperduti, A. 2003. Discretizing continuous attributes in AdaBoost for text categorization. ADVANCES IN INFORMATION RETRIEVAL 2633: 320-334. LECTURE NOTES IN COMPUTER SCIENCE
Nasukawa, T; Nagano, T. 2001. Text analysis and knowledge mining system. IBM SYSTEMS JOURNAL 40 (4): 967-984.
Nenadic, G; Mima, H; Spasic, I; Ananiadou, S; Tsujii, J. 2002. Terminology-driven literature mining and knowledge acquisition in biomedicine. INTERNATIONAL JOURNAL OF MEDICAL INFORMATICS 67 (1-3): 33-48.
Nenadic, G; Spasic, I; Ananiadou, S. 2003. Terminology-driven mining of biomedical literature. BIOINFORMATICS 19 (8): 938-943.
Newby, GB. 2001. Empirical study of a 3D visualization for information retrieval tasks. JOURNAL OF INTELLIGENT INFORMATION SYSTEMS 18 (1): 31-53.
Nguyen, CD; Dung, TA; Cao, TH. 2005. Text classification for DAG-structured categories. ADVANCES IN KNOWLEDGE DISCOVERY AND DATA MINING, PROCEEDINGS 3518: 290-300. LECTURE NOTES IN ARTIFICIAL INTELLIGENCE
Nigam, K; McCallum, AK; Thrun, S; Mitchell, T. 2000. Text classification from labeled and unlabeled documents using EM. MACHINE LEARNING 39 (2-3): 103-134.

Niimi, A; Konishi, O. 2003. Data mining for distributed Databases with multiagents. KNOWLEDGE-BASED INTELLIGNET INFORMATION AND ENGINEERING SYSTEMS, PT 2, PROCEEDINGS 2774: 1412-1418. LECTURE NOTES IN ARTIFICIAL INTELLIGENCE
Niimi, A; Konishi, O. 2004. Extension of multiagent data mining for distributed databases. KNOWLEDGE-BASED INTELLIGENT INFORMATION AND ENGINEERING SYSTEMS, PT 3, PROCEEDINGS 3215: 780-787. LECTURE NOTES IN ARTIFICIAL INTELLIGENCE
Niimi, A; Noji, H; Konishi, O. 2005. Distributed web integration with multiagent data mining. KNOWLEDGE-BASED INTELLIGENT INFORMATION AND ENGINEERING SYSTEMS, PT 3, PROCEEDINGS 3683: 513-519. LECTURE NOTES IN ARTIFICIAL INTELLIGENCE
Nomoto, T; Matsumoto, Y. 2003. The diversity-based approach to open-domain text summarization. INFORMATION PROCESSING & MANAGEMENT 39 (3): 363-389.
Novovicova, J; Malik, A. 2003. Application of multinomial mixture model to text classification. PATTERN RECOGNITION AND IMAGE ANALYSIS, PROCEEDINGS 2652: 646-653. LECTURE NOTES IN COMPUTER SCIENCE
Novovicova, J; Malik, A. 2004. Text document classification based on mixture models. KYBERNETIKA 40 (3): 293-304.
Novovicova, J; Malik, A; Pudil, P. 2004. Feature selection using improved mutual information for text classification. STRUCTURAL, SYNTACTIC, AND STATISTICAL PATTERN RECOGNITION, PROCEEDINGS 3138: 1010-1017. LECTURE NOTES IN COMPUTER SCIENCE
Ohsumi, N; Yasuda, A. 2004. Reviewing textual data mining in japan. SOCIOLOGICAL THEORY AND METHODS 19 (2): 135-159.
Oliveira, S; Seok, SC. 2005. A multi-level approach for document clustering. COMPUTATIONAL SCIENCE - ICCS 2005, PT 1, PROCEEDINGS 3514: 204-211. LECTURE NOTES IN COMPUTER SCIENCE
Oliver, DE; Bhalotia, G; Schwartz, AS; Altman, RB; Hearst, MA. 2004. Tools for loading MEDLINE into a local relational database. BMC BIOINFORMATICS 5: art. no.-146.
Omelayenko, B. 2002. Integrating vocabularies: Discovering and representing vocabulary maps. SEMANTIC WEB - ISWC 2002 2342: 206-220. LECTURE NOTES IN COMPUTER SCIENCE

Ong, TH; Chen, HC; Sung, WK; Zhu, B. 2005. Newsmap: a knowledge map for online news. DECISION SUPPORT SYSTEMS 39 (4): 583-597.
Onodera, N. 2001. A bibliometric study on chemical information and computer sciences focusing on literature of JCICS. JOURNAL OF CHEMICAL INFORMATION AND COMPUTER SCIENCES 41 (4): 878-888.
Orengo, VM; Huyck, C. 2003. Portuguese-English experiments using latent semantic indexing. ADVANCES IN CROSS-LANGUAGE INFORMATION RETRIEVAL 2785: 147-154. LECTURE NOTES IN COMPUTER SCIENCE
OTTAVIANI, JS. 1994. THE FRACTAL NATURE OF RELEVANCE - A HYPOTHESIS. JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE 45 (4): 263-272.
Oudshoff, AM; Bosloper, IE; Klos, TB; Spaanenburg, L. 2003. Knowledge discovery in virtual community texts: Clustering virtual communities. JOURNAL OF INTELLIGENT & FUZZY SYSTEMS 14 (1): 13-24.
Ozono, T; Shintani, T; Ito, T; Hasegawa, T. 2004. A feature selection for text categorization on research support system papits. PRICAI 2004: TRENDS IN ARTIFICIAL INTELLIGENCE, PROCEEDINGS 3157: 524-533. LECTURE NOTES IN ARTIFICIAL INTELLIGENCE
Pal, U; Chaudhuri, BB. 2001. Machine-printed and hand-written text lines identification. PATTERN RECOGNITION LETTERS 22 (3-4): 431-441.
Palotai, Z; Gabor, B; Lorincz, A. 2005. Adaptive highlighting of links to assist surfing on the Internet. INTERNATIONAL JOURNAL OF INFORMATION TECHNOLOGY & DECISION MAKING 4 (1): 117-139.
Pan, H; Zuo, L; Choudhary, V; Zhang, Z; Leow, SH; Chong, FT; Huang, YL; Ong, VWS; Mohanty, B; Tan, SL; Krishnan, SPT; Bajic, VB. 2004. Dragon TF Association Miner: a system for exploring transcription factor associations through text-mining. NUCLEIC ACIDS RESEARCH 32: W230-W234, Suppl. 2.
Pan, LY; Song, H; Ma, FY. 2004. A macrocommittees method of combining multistrategy classifiers for heterogeneous ontology matching. ADVANCES IN WEB-AGE INFORMATION MANAGEMENT: PROCEEDINGS 3129: 672-677. LECTURE NOTES IN COMPUTER SCIENCE
Papadimitriou, CH; Raghavan, P; Tamaki, H; Vempala, S. 2000. Latent semantic indexing: A probabilistic analysis. JOURNAL OF COMPUTER AND SYSTEM SCIENCES 61 (2): 217-235.
Pappuswamy, U; Bhembe, D; Jordan, PW; VanLehn, K. 2005. A supervised clustering method for text classification. COMPUTATIONAL

LINGUISTICS AND INTELLIGENT TEXT PROCESSING 3406: 704-714. LECTURE NOTES IN COMPUTER SCIENCE
Paradis, RO; Nie, JY. 2005. Filtering contents with bigrams and named entities to improve text classification. INFORMATION RETRIEVAL TECHNOLOGY, PROCEEDINGS 3689: 135-146. LECTURE NOTES IN COMPUTER SCIENCE
Park, BK; Han, H; Song, IY. 2005. XML-OLAP: A multidimensional analysis framework for XML warehouses. DATA WAREHOUSING AND KNOWLEDGE DISCOVERY, PROCEEDINGS 3589: 32-42. LECTURE NOTES IN COMPUTER SCIENCE
Park, H; Jeon, M; Ben Rosen, J. 2003. Lower dimensional representation of text data based on centroids and least squares. BIT 43 (2): 427-448.
Park, HS; Kim, MK; Choi, EJ; Seol, YJ. 2005. Text mining from categorized stem cell documents to infer developmental stage-specific expression and regulation patterns of stem cells. NATURAL LANGUAGE PROCESSING AND INFORMATION SYSTEMS, PROCEEDINGS 3513: 353-356. LECTURE NOTES IN COMPUTER SCIENCE
Park, J; Lee, C; Park, JC. 2005. Information visualization with text data mining for knowledge discovery tools in bioinformatics. ON THE CONVERGENCE OF BIO-INFORMATION-, ENVIRONMENTAL-, ENERGY-, SPACE- AND NANO-TECHNOLOGIES, PTS 1 AND 2 277-279: 259-265.
KEY ENGINEERING MATERIALS
Park, SB; Zhang, BT. 2004. Co-trained support vector machines for large scale unstructured document classification using unlabeled data and syntactic information. INFORMATION PROCESSING & MANAGEMENT 40 (3): 421-439.
Patman, F; Thompson, P. 2003. Names: A new frontier in text mining. INTELLIGENCE AND SECURITY INFORMATICS, PROCEEDINGS 2665: 27-38. LECTURE NOTES IN COMPUTER SCIENCE
Peng, FC; Schuurmans, D. 2003. Combining naive Bayes and n-gram language models for text classification. ADVANCES IN INFORMATION RETRIEVAL 2633: 335-350. LECTURE NOTES IN COMPUTER SCIENCE
Peng, FC; Schuurmans, D; Wang, SJ. 2004. Augmenting naive Bayes classifiers with statistical language models. INFORMATION RETRIEVAL 7 (3-4): 317-345.
Percannella, G; Sorrentino, D; Vento, M. 2005. Automatic indexing of news videos through text classification techniques. PATTERN RECOGNITION

AND IMAGE ANALYSIS, PT 2, PROCEEDINGS 3687: 512-521. LECTURE NOTES IN COMPUTER SCIENCE
Perez-Sancho, C; Inesta, JM; Calera-Rubio, J. 2005. A text categorization approach for music style recognition. PATTERN RECOGNITION AND IMAGE ANALYSIS, PT 2, PROCEEDINGS 3523: 649-657. LECTURE NOTES IN COMPUTER SCIENCE
Perrin, P; Petry, FE. 2003. Extraction and representation of contextual information for knowledge discovery in texts. INFORMATION SCIENCES 151: 125-152.
Persidis, A; Deftereos, S; Persidis, A. 2004. Systems literature analysis. PHARMACOGENOMICS 5 (7): 943-947.
Peters, C; Koster, CHA. 2002. Uncertainty-based noise reduction and term selection in text categorization. ADVANCES IN INFORMATION REFTRIEVAL 2291: 248-267. LECTURE NOTES IN COMPUTER SCIENCE
Peters, CMEE; Koster, CHA. 2003. Uncertainty and term selection in text categorization. INTERNATIONAL JOURNAL OF UNCERTAINTY FUZZINESS AND KNOWLEDGE-BASED SYSTEMS 11 (1): 115-137.
Petridis, V; Kaburlasos, VG. 2001. Clustering and classification in structured data domains using Fuzzy Lattice Neurocomputing (FLN). IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING 13 (2): 245-260.
Pierrakos, D; Paliouras, G; Papatheodorou, C; Karkaletsis, V; Dikaiakos, M. 2004. Web Community Directories: A new approach to Web Personalization. WEB MINING: FROM WEB TO SEMANTIC WEB 3209: 113-129. LECTURE NOTES IN COMPUTER SCIENCE
Pierre, JM. 2002. Mining knowledge from text collections using automatically generated metadata. PRACTICAL ASPECTS OF KNOWLEDGE MANAGEMENT 2569: 537-548. LECTURE NOTES IN ARTIFICIAL INTELLIGENCE
Pongpinigpinyo, S; Rivepiboon, W. 2005. Word sense disambiguation of Thai language with unsupervised learning. KNOWLEDGE-BASED INTELLIGENT INFORMATION AND ENGINEERING SYSTEMS, PT 1, PROCEEDINGS 3681: 1275-1283. LECTURE NOTES IN ARTIFICIAL INTELLIGENCE
Pons-Porrata, A; Berlanga-Llavori, R; Ruiz-Shulcloper, J. 2003. Building a hierarchy of events and topics for newspaper digital libraries. ADVANCES IN INFORMATION RETRIEVAL 2633: 588-596. LECTURE NOTES IN COMPUTER SCIENCE

Pons-Porrata, A; Ruiz-Shulcloper, J; Berlanga-Llavori, R. 2003. A method for the automatic summarization of topic-based clusters of documents. PROGRESS IN PATTERN RECOGNITION, SPEECH AND IMAGE ANALYSIS 2905: 596-603. LECTURE NOTES IN COMPUTER SCIENCE
Porter, AL; Kongthon, A; Lui, JC. 2002. Research profiling: Improving the literature review. SCIENTOMETRICS 53 (3): 351-370.
Prasad, PC; Arunkumar, S. 2004. From short-term memory to semantics-a computational model. NEURAL COMPUTING & APPLICATIONS 13 (2): 157-167.
Price, L; Thelwall, M. 2005. The clustering power of low frequency words in academic webs. JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE AND TECHNOLOGY 56 (8): 883-888.
Price, RJ; Zukas, AE. 2005. Application of latent semantic indexing to processing of noisy text. INTELLIGENCE AND SECURITY INFORMATICS, PROCEEDINGS 3495: 602-603. LECTURE NOTES IN COMPUTER SCIENCE
Pullwitt, D. 2002. Integrating contextual information to enhance SOM-based text document clustering. NEURAL NETWORKS 15 (8-9): 1099-1106.
Qian, TY; Wang, YZ; Long, H; Feng, JL. 2005. 2-PS based associative text classification. DATA WAREHOUSING AND KNOWLEDGE DISCOVERY, PROCEEDINGS 3589: 378-387. LECTURE NOTES IN COMPUTER SCIENCE
Qian, WN; Zhang, L; Liang, YQ; Qian, HL; Jin, W. 2000. A two-level method for clustering DTDs. WEB-AGE INFORMATION MANAGEMENT, PROCEEDINGS 1846: 41-52. LECTURE NOTES IN COMPUTER SCIENCE
Qiang, W; Wang, XL; Yi, G. 2005. A study of semi-discrete matrix decomposition for LSI in automated text categorization. NATURAL LANGUAGE PROCESSING - IJCNLP 2004 3248: 606-615. LECTURE NOTES IN COMPUTER SCIENCE
Quan, TT; Hui, SC; Fong, A. 2003. Mining multiple clustering data for knowledge discovery. DISCOVERY SCIENCE, PROCEEDINGS 2843: 452-459. LECTURE NOTES IN ARTIFICIAL INTELLIGENCE
Raez, AM; Lopez, LAU; Lopez, LAU; Steinberger, R. 2004. Adaptive selection of base classifiers in one-against-all learning for large multi-labeled collections. ADVANCES IN NATURAL LANGUAGE PROCESSING 3230: 1-12. LECTURE NOTES IN ARTIFICIAL INTELLIGENCE

Raj, PCR; Raman, S. 2005. A phrase grammar-based conceptual indexing paradigm. APPLIED ARTIFICIAL INTELLIGENCE 19 (6): 559-599.
Rauber, A; Merkl, D. 1999. The SOMLib digital library system. RESEARCH AND ADVANCED TECHNOLOGY FOR DIGITAL LIBRARIES, PROCEEDINGS 1696: 323-342. LECTURE NOTES IN COMPUTER SCIENCE
Rauber, A; Merkl, D. 2003. Text mining in the SOMLib Digital Library System: The representation of topics and genres. APPLIED INTELLIGENCE 18 (3): 271-293.
Rehel, S; Mineau, GW. 2005. Vocabulary completion through word cooccurrence analysis using unlabeled documents for text categorization. ADVANCES IN ARTIFICIAL INTELLIGENCE, PROCEEDINGS 3501: 377-388. LECTURE NOTES IN COMPUTER SCIENCE
Reinberger, ML; Daelemans, W. 2003. Is shallow parsing useful for unsupervised learning of semantic clusters?. COMPUTATIONAL LINGUISTICS AND INTELLIGENT TEXT PROCESSING, PROCEEDINGS 2588: 304-313. LECTURE NOTES IN COMPUTER SCIENCE
Reinberger, ML; Spyns, P; Daelemans, W; Meersman, R. 2003. Mining for lexons: Applying unsupervised learning methods to create ontology bases. ON THE MOVE TO MEANINGFUL INTERNET SYSTEMS 2003: COOPIS, DOA, AND ODBASE 2888: 803-819. LECTURE NOTES IN COMPUTER SCIENCE
Reinberger, ML; Spyns, P; Pretorius, AJ; Daelemans, W. 2004. Automatic initiation of an ontology. ON THE MOVE TO MEANINGFUL INTERNET SYSTEMS 2004: COOPIS, DOA, AND ODBASE, PT 1, PROCEEDINGS 3290: 600-617. LECTURE NOTES IN COMPUTER SCIENCE
Remeikis, N; Skucas, I; Melninkaite, V. 2004. Text categorization using neural networks initialized with decision trees. INFORMATICA 15 (4): 551-564.
RILOFF, E; LEHNERT, W. 1994. INFORMATION EXTRACTION AS A BASIS FOR HIGH-PRECISION TEXT CLASSIFICATION. ACM TRANSACTIONS ON INFORMATION SYSTEMS 12 (3): 296-333.
Rizzo, R; Munna, EG. 2000. A neural network tool to organize large document sets. ARTIFICIAL INTELLIGENCE: METHODOLOGY, SYSTEMS, APPLICATIONS, PROCEEDINGS 1904: 301-309. LECTURE NOTES IN ARTIFICIAL INTELLIGENCE
Rogozan, A; Neveol, A; Darmoni, SJ. 2003. Text categorization prior to indexing for the CISMEF health catalogue. ARTIFICIAL INTELLIGENCE

IN MEDICINE, PROCEEDINGS 2780: 81-85. LECTURE NOTES IN ARTIFICIAL INTELLIGENCE
Romero, E; Marquez, L; Carreras, X. 2004. Margin maximization with feed-forward neural networks: a comparative study with SVM and AdaBoost. NEUROCOMPUTING 57: 313-344.
Rosso, P; Molina, A; Pla, F; Jimenez, D; Vidal, V. 2004. Information retrieval and text categorization with semantic indexing. COMPUTATIONAL LINGUISTICS AND INTELLIGENT TEXT PROCESSING 2945: 596-600. LECTURE NOTES IN COMPUTER SCIENCE
Roussinov, D; Zhao, JL. 2003. Automatic discovery of similarity relationships through Web mining. DECISION SUPPORT SYSTEMS 35 (1): 149-166.
Roussinov, DG; Chen, HC. 1999. Document clustering for electronic meetings: an experimental comparison of two techniques. DECISION SUPPORT SYSTEMS 27 (1-2): 67-79.
Roussinov, DG; Chen, HC. 2001. Information navigation on the web by clustering and summarizing query results. INFORMATION PROCESSING & MANAGEMENT 37 (6): 789-816.
Ruch, P; Baud, R; Geissbuhler, A. 2003. Learning-free text categorization. ARTIFICIAL INTELLIGENCE IN MEDICINE, PROCEEDINGS 2780: 199-208. LECTURE NOTES IN ARTIFICIAL INTELLIGENCE
Ruiz, ME; Srinivasan, P. 2002. Hierarchical text categorization using neural networks. INFORMATION RETRIEVAL 5 (1): 87-118.
Ruiz, ME; Srinivasan, P. 2003. Hybrid hierarchical classifiers for categorization of medical documents. ASIST 2003: PROCEEDINGS OF THE 66TH ASIST ANNUAL MEETING, VOL 40, 2003 40: 65-70. PROCEEDINGS OF THE ASIST ANNUAL MEETING
Russell, B; Yin, HJ; Allinson, NM. 2002. Document clustering using the 1+1 dimensional Self-Organising Map. INTELLIGENT DATA ENGINEERING AND AUTOMATED LEARNING - IDEAL 2002 2412: 154-160. LECTURE NOTES IN COMPUTER SCIENCE
Rzhetsky, A; Iossifov, I; Koike, T; Krauthammer, M; Kra, P; Morris, M; Yu, H; Duboue, PA; Weng, WB; Wilbur, WJ; Hatzivassiloglou, V; Friedman, C. 2004. GeneWays: a system for extracting, analyzing, visualizing, and integrating molecular pathway data. JOURNAL OF BIOMEDICAL INFORMATICS 37 (1): 43-53.
Sadakane, K; Imai, H. 2001. Fast algorithms for k-word proximity search. IEICE TRANSACTIONS ON FUNDAMENTALS OF ELECTRONICS

COMMUNICATIONS AND COMPUTER SCIENCES E84A (9): 2311-2318.
Saidi, AS. 2005. Using CLP to characterise linguistic lattice boundaries in a text mining process. LOGIC PROGRAMMING, PROCEEDINGS 3668: 418-420. LECTURE NOTES IN COMPUTER SCIENCE
Sakakibara, Y; Misue, K; Koshiba, T. 1996. A machine learning approach to knowledge acquisitions from text databases. INTERNATIONAL JOURNAL OF HUMAN-COMPUTER INTERACTION 8 (3): 309-324.
Sakkis, G; Androutsopoulos, I; Paliouras, G; Karkaletsis, V; Spyropoulos, CD; Stamatopoulos, P. 2003. A memory-based approach to anti-spam filtering for mailing lists. INFORMATION RETRIEVAL 6 (1): 49-73.
Sakurai, S; Ichimura, Y; Suyama, A; Orihara, R. 2002. Acquisition of a knowledge dictionary from training examples including multiple values. FOUNDATIONS OF INTELLIGENT SYSTEMS, PROCEEDINGS 2366: 103-113. LECTURE NOTES IN ARTIFICIAL INTELLIGENCE
Sakurai, S; Suyama, A. 2005. An e-mail analysis method based on text mining techniques. APPLIED SOFT COMPUTING 6 (1): 62-71.
Sanchez, VD. 2003. Advanced support vector machines and kernel methods. NEUROCOMPUTING 55 (1-2): 5-20.
SanJuan, E; Dowdall, J; Ibekwe-SanJuan, F; Rinaldi, F. 2005. A symbolic approach to automatic multiword term structuring. COMPUTER SPEECH AND LANGUAGE 19 (4): 524-542.
Sanz, I; Perez, JM; Berlanga, R; Aramburu, MJ. 2003. XML schemata inference and evolution. DATABASE AND EXPERT SYSTEMS APPLICATIONS, PROCEEDINGS 2736: 109-118. LECTURE NOTES IN COMPUTER SCIENCE
Saravanan, M; Raj, PCR; Raman, S. 2003. Summarization and categorization of text data in high-level data cleaning for information retrieval. APPLIED ARTIFICIAL INTELLIGENCE 17 (5-6): 461-474.
Sarkar, K; Bandyopadhyay, S. 2005. Generating headline summary from a document set. COMPUTATIONAL LINGUISTICS AND INTELLIGENT TEXT PROCESSING 3406: 649-652. LECTURE NOTES IN COMPUTER SCIENCE
Sauban, M; Pfahringer, B. 2003. Text categorisation using document profiling. KNOWLEDGE DISCOVERY IN DATABASES: PKDD 2003, PROCEEDINGS 2838: 411-422. LECTURE NOTES IN ARTIFICIAL INTELLIGENCE
Scarinci, RG; Wives, LK; Loh, S; Zabenedetti, C; de Oliveira, JPM. 2003. Managing unstructured E-commerce information. ADVANCED

CONCEPTUAL MODELING TECHNIQUES 2784: 414-426. LECTURE NOTES IN COMPUTER SCIENCE
Schaal, M; Muller, RM; Brunzel, M; Spiliopoulou, M. 2005. RELFIN - Topic discovery for ontology enhancement and annotation. SEMANTIC WEB: RESEARCH AND APPLICATIONS, PROCEEDINGS 3532: 608-622. LECTURE NOTES IN COMPUTER SCIENCE
Schapire, RE; Singer, Y. 2000. BoosTexter: A boosting-based system for text categorization. MACHINE LEARNING 39 (2-3): 135-168.
Scharl, A; Bauer, C. 2004. Mining large samples of web-based corpora. KNOWLEDGE-BASED SYSTEMS 17 (5-6): 229-233.
Schenker, A; Bunke, H; Last, M; Kandel, A. 2004. A graph-based framework for Web document mining. DOCUMENT ANALYSIS SYSTEMS VI, PROCEEDINGS 3163: 401-412. LECTURE NOTES IN COMPUTER SCIENCE
Schenker, A; Bunke, H; Last, M; Kandel, A. 2004. Building graph-based classifier ensembles by random node selection. MULTIPLE CLASSIFIER SYSTEMS, PROCEEDINGS 3077: 214-222. LECTURE NOTES IN COMPUTER SCIENCE
Schenker, A; Last, M; Bunke, H; Kandel, A. 2003. Graph representations for Web document clustering. PATTERN RECOGNITION AND IMAGE ANALYSIS, PROCEEDINGS 2652: 935-942. LECTURE NOTES IN COMPUTER SCIENCE
Schenker, A; Last, M; Bunke, H; Kandel, A. 2004. Classification of web documents using graph matching. INTERNATIONAL JOURNAL OF PATTERN RECOGNITION AND ARTIFICIAL INTELLIGENCE 18 (3): 475-496.
Schenker, A; Last, M; Bunke, H; Kandell, A. 2004. Comparison of algorithms for web document clustering using graph representations of data. STRUCTURAL, SYNTACTIC, AND STATISTICAL PATTERN RECOGNITION, PROCEEDINGS 3138: 190-197. LECTURE NOTES IN COMPUTER SCIENCE
Scherf, M; Epple, A; Werner, T. 2005. The next generation of literature analysis: Integration of genomic analysis into text mining. BRIEFINGS IN BIOINFORMATICS 6 (3): 287-297.
Schijvenaars, BJA; Mons, B; Weeber, M; Schuemie, MJ; van Mulligen, EM; Wain, HM; Kors, JA. 2005. Thesaurus-based disambiguation of gene symbols. BMC BIOINFORMATICS 6: art. no.-149.
Schneider, KM. 2004. On word frequency information and negative evidence in Naive Bayes text classification. ADVANCES IN NATURAL

LANGUAGE PROCESSING 3230: 474-485. LECTURE NOTES IN ARTIFICIAL INTELLIGENCE
Schneider, KM. 2005. Learning to filter junk e-mail from positive and unlabeled examples. NATURAL LANGUAGE PROCESSING - IJCNLP 2004 3248: 426-435. LECTURE NOTES IN COMPUTER SCIENCE
Schneider, KM. 2005. Techniques for improving the performance of naive Bayes for text classification. COMPUTATIONAL LINGUISTICS AND INTELLIGENT TEXT PROCESSING 3406: 682-693. LECTURE NOTES IN COMPUTER SCIENCE
Schneider, KM. 2005. Weighted average pointwise mutual information for feature selection in text categorization. KNOWLEDGE DISCOVERY IN DATABASES: PKDD 2005 3721: 252-263. LECTURE NOTES IN ARTIFICIAL INTELLIGENCE
Schuemie, MJ; Kors, JA; Mons, B. 2005. Word sense disambiguation in the biomedical domain: An overview. JOURNAL OF COMPUTATIONAL BIOLOGY 12 (5): 554-565.
Sclaroff, S; La Cascia, M; Sethi, S; Taycher, L. 1999. Unifying textual and visual cues for content-based image retrieval on the World Wide Web. COMPUTER VISION AND IMAGE UNDERSTANDING 75 (1-2): 86-98.
Seki, K; Mostafa, J. 2005. A hybrid approach to protein name identification in biomedical texts. INFORMATION PROCESSING & MANAGEMENT 41 (4): 723-743.
Selamat, A; Omatu, S. 2004. Web page feature selection and classification using neural networks. INFORMATION SCIENCES 158: 69-88.
Settles, B. 2005. ABNER: an open source tool for automatically tagging genes, proteins and other entity names in text. BIOINFORMATICS 21 (14): 3191-3192.
Sever, H; Bolat, Z; Raghavan, VV. 2004. Use of preference relation for text categorization. ROUGH SETS AND CURRENT TRENDS IN COMPUTING 3066: 708-713. LECTURE NOTES IN ARTIFICIAL INTELLIGENCE
Sever, H; Gorur, A; Tolun, MR. 2003. Text categorization with ILA. COMPUTER AND INFORMATION SCIENCES - ISCIS 2003 2869: 300-307. LECTURE NOTES IN COMPUTER SCIENCE
Sevillano, X; Alias, F; Socoro, JC. 2004. Reliability in ICA-based text classification. INDEPENDENT COMPONENT ANALYSIS AND BLIND SIGNAL SEPARATION 3195: 1213-1220. LECTURE NOTES IN COMPUTER SCIENCE

Shahnaz, F; Berry, MW; Pauca, VP; Plemmons, RJ. 2006. Document clustering using nonnegative matrix factorization/. INFORMATION PROCESSING & MANAGEMENT 42 (2): 373-386.
Shanahan, JG; Roma, N. 2003. Improving SVM text classification performance through threshold adjustment. MACHINE LEARNING: ECML 2003 2837: 361-372. LECTURE NOTES IN ARTIFICIAL INTELLIGENCE
Shatkay, H. 2005. Hairpins in bookstacks: Information retrieval from biomedical text. BRIEFINGS IN BIOINFORMATICS 6 (3): 222-238.
SHAW, RJ; WILLETT, P. 1993. ON THE NONRANDOM NATURE OF NEAREST-NEIGHBOR DOCUMENT CLUSTERS. INFORMATION PROCESSING & MANAGEMENT 29 (4): 449-452.
Shi, M; Edwin, DS; Menon, R; Shen, LX; Lim, JYK; Loh, HT; Keerthi, SS; Ong, CJ. 2003. A machine learning approach for the curation of biomedical literature. ADVANCES IN INFORMATION RETRIEVAL 2633: 597-604. LECTURE NOTES IN COMPUTER SCIENCE
Shima, K; Todoriki, M; Suzuki, A. 2004. SVM-based feature selection of latent semantic features. PATTERN RECOGNITION LETTERS 25 (9): 1051-1057.
Shimazu, K; Momma, A; Furukawa, K. 2003. Discovering exceptional information from customer inquiry by association rule miner. DISCOVERY SCIENCE, PROCEEDINGS 2843: 269-282. LECTURE NOTES IN ARTIFICIAL INTELLIGENCE
Shimbo, M; Yamasaki, T; Matsumoto, Y. 2005. Sentence role identification in medline abstracts: Training classifier with structured abstracts. ACTIVE MINING 3430: 236-254. LECTURE NOTES IN ARTIFICIAL INTELLIGENCE
Sidhom, S; Hassoun, M. 2002. Morpho-syntactic parsing for a text mining environment: An NP recognition model for knowledge visualization and information retrieval. KNOWLEDGE ORGANIZATION 29 (3): 171-180.
Siefkes, C; Assis, F; Chhabra, S; Yerazunis, WS. 2004. Combining Winnow and orthogonal sparse bigrams for incremental spam filtering. KNOWLEDGE DISCOVERY IN DATABASES: PKDD 2004, PROCEEDINGS 3202: 410-421. LECTURE NOTES IN ARTIFICIAL INTELLIGENCE
Silverstein, C; Brin, S; Motwani, R. 1998. Beyond market baskets: Generalizing association rules to dependence rules. DATA MINING AND KNOWLEDGE DISCOVERY 2 (1): 39-68.

Silverstein, C; Brin, S; Motwani, R; Ullman, J. 2000. Scalable techniques for mining causal structures. DATA MINING AND KNOWLEDGE DISCOVERY 4 (2-3): 163-192.
Singh, P; Bhimavarapu, R; Davulcu, H; Baral, C; Kim, S; Liu, H; Bittner, M; Ramakrishnan, IV. 2005. BioLog: A browser based collaboration and resource navigation assistant for biomedical researchers. DATA INTEGRATION IN THE LIFE SCIENCES, PROCEEDINGS 3615: 19-30. LECTURE NOTES IN COMPUTER SCIENCE
Singh, P; Bhimavarapu, R; Davulcu, H; Baral, C; Kim, S; Liu, H; Bittner, M; Ramakrishnan, IV. 2005. BioLog: A browser based collaboration and resource navigation assistant for biomedical researchers. DATA INTEGRATION IN THE LIFE SCIENCES, PROCEEDINGS 3615: 19-30. LECTURE NOTES IN COMPUTER SCIENCE
Singh, SB; Hull, RD; Fluder, EM. 2003. Text Influenced Molecular Indexing (TIMI): A literature database mining approach that handles text and chemistry. JOURNAL OF CHEMICAL INFORMATION AND COMPUTER SCIENCES 43 (3): 743-752.
Singh, SB; Sheridan, RP; Fluder, EM; Hull, RD. 2001. Mining the chemical quarry with joint chemical probes: An application of latent semantic structure indexing (LaSSI) and TOPOSIM (dice) to chemical database mining. JOURNAL OF MEDICINAL CHEMISTRY 44 (10): 1564-1575.
Sinka, MP; Corne, DW. 2005. The BankSearch web document dataset: investigating unsupervised clustering and category similarity. JOURNAL OF NETWORK AND COMPUTER APPLICATIONS 28 (2): 129-146.
Skarmeta, AG; Bensaid, A; Tazi, N. 2000. Data mining for text categorization with semi-supervised agglomerative hierarchical clustering. INTERNATIONAL JOURNAL OF INTELLIGENT SYSTEMS 15 (7): 633-646.
Slaney, M; Subrahmonia, J; Maglio, P. 2003. Modeling multitasking users. USER MODELING 2003, PROCEEDINGS 2702: 188-197. LECTURE NOTES IN ARTIFICIAL INTELLIGENCE
Smalheiser, NR. 2001. Predicting emerging technologies with the aid of text-based data mining: the micro approach. TECHNOVATION 21 (10): 689-693.
Smalheiser, NR; Palmer, CL; Swanson, DR; Srinivasan, P; Hearst, M. 2003. Literature-based discovery: New trends and techniques. ASIST 2003: PROCEEDINGS OF THE 66TH ASIST ANNUAL MEETING, VOL 40, 2003 40: 497-497. PROCEEDINGS OF THE ASIST ANNUAL MEETING
SMALHEISER, NR; SWANSON, DR. 1994. ASSESSING A GAP IN THE BIOMEDICAL LITERATURE - MAGNESIUM-DEFICIENCY AND

NEUROLOGIC DISEASE. NEUROSCIENCE RESEARCH COMMUNICATIONS 15 (1): 1-9.
Smalheiser, NR; Swanson, DR. 1996. Indomethacin and Alzheimer's disease. NEUROLOGY 46 (2): 583-583.
Smalheiser, NR; Swanson, DR. 1996. Linking estrogen to Alzheimer's disease: An informatics approach. NEUROLOGY 47 (3): 809-810.
Smalheiser, NR; Swanson, DR. 1998. Using ARROWSMITH: a computer-assisted approach to formulating and assessing scientific hypotheses. COMPUTER METHODS AND PROGRAMS IN BIOMEDICINE 57 (3): 149-153.
Smirnov, A; Pashkin, M; Chilov, N; Levashova, T; Krizhanovsky, A; Kashevnik, A. 2005. Ontology-based users and requests clustering in customer service management system. AUTONOMOUS INTELLIGENT SYSTEMS: AGENTS AND DATA MINING, PROCEEDINGS 3505: 231-246. LECTURE NOTES IN COMPUTER SCIENCE
Smith, B; Kohler, J; Kumar, A. 2004. On the application of formal principles to life science data: a case study in the gene ontology. DATA INTEGRATION IN THE LIFE SCIENCES, PROCEEDINGS 2994: 79-94. LECTURE NOTES IN BIOINFORMATICS
Soboroff, IM; Nicholas, CK. 2002. Related, but not relevant: Content-based collaborative filtering in TREC-8. INFORMATION RETRIEVAL 5 (2-3): 189-208.
Sohler, F; Hanisch, D; Zimmer, R. 2004. New methods for joint analysis of biological networks and expression data. BIOINFORMATICS 20 (10): 1517-1521.
Song, FX; Liu, SH; Yang, JY. 2005. A comparative study on text representation schemes in text categorization. PATTERN ANALYSIS AND APPLICATIONS 8 (1-2): 199-209.
Song, XD; Tseng, BL; Lin, CY; Sun, MT. 2005. ExpertiseNet: Relational and evolutionary expert modeling. USER MODELING 2005, PROCEEDINGS 3538: 99-108. LECTURE NOTES IN ARTIFICIAL INTELLIGENCE
Soonthornphisaj, N; Kijsirikul, B. 2004. Iterative cross-training: An algorithm for learning from unlabeled Web pages. INTERNATIONAL JOURNAL OF INTELLIGENT SYSTEMS 19 (1-2): 131-147.
Soucy, P; Mineau, GW. 2003. Feature selection strategies for text categorization. ADVANCES IN ARTIFICIAL INTELLIGENCE, PROCEEDINGS 2671: 505-509. LECTURE NOTES IN ARTIFICIAL INTELLIGENCE

Spangler, S; Kreulen, JT; Lessler, J. 2003. Generating and browsing multiple taxonomies over a document collection. JOURNAL OF MANAGEMENT INFORMATION SYSTEMS 19 (4): 191-212.
Spasic, I; Ananiadou, S; McNaught, J; Kumar, A. 2005. Text mining and ontologies in biomedicine: Making sense of raw text. BRIEFINGS IN BIOINFORMATICS 6 (3): 239-251.
Spiliopoulou, M. 1999. Data mining for the web. PRINCIPLES OF DATA MINING AND KNOWLEDGE DISCOVERY 1704: 588-589. LECTURE NOTES IN ARTIFICIAL INTELLIGENCE
Srinivasan, P. 2001. MeSHmap: A text mining tool for MEDLINE. JOURNAL OF THE AMERICAN MEDICAL INFORMATICS ASSOCIATION: 642-646, Suppl. S.
Srinivasan, P. 2004. Text mining: Generating hypotheses from MEDLINE. JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE AND TECHNOLOGY 55 (5): 396-413.
Stamatatos, E; Kokkinakis, G; Fakotakis, N. 2000. Automatic text categorization in terms of genre and author. COMPUTATIONAL LINGUISTICS 26 (4): 471-495.
Stegmann, J; Grohmann, G. 2003. Hypothesis generation guided by co-word clustering. SCIENTOMETRICS 56 (1): 111-135.
Stein, B; Eissen, SMZ. 2003. Automatic document categorization - Interpreting the perfomance of clustering algorithms. KI 2003: ADVANCES IN ARTIFICIAL INTELLIGENCE 2821: 254-266. LECTURE NOTES IN ARTIFICIAL INTELLIGENCE
Stensmo, M. 2002. A scalable and efficient probabilistic information retrieval and text mining system. ARTIFICIAL NEURAL NETWORKS - ICANN 2002 2415: 643-648. LECTURE NOTES IN COMPUTER SCIENCE
Story, RE. 1996. An explanation of the effectiveness of latent semantic indexing by means of a Bayesian regression model. INFORMATION PROCESSING & MANAGEMENT 32 (3): 329-344.
Strehl, A; Ghosh, J. 2003. Relationship-based clustering and visualization for high-dimensional data mining. INFORMS JOURNAL ON COMPUTING 15 (2): 208-230.
Su, Z; Zhang, L; Pan, Y. 2003. Document clustering based on vector quantization and growing-cell structure. DEVELOPMENTS IN APPLIED ARTIFICIAL INTELLIGENCE 2718: 326-336. LECTURE NOTES IN ARTIFICIAL INTELLIGENCE
Subasic, P; Huettner, A. 2001. Affect analysis of text using fuzzy semantic typing. IEEE TRANSACTIONS ON FUZZY SYSTEMS 9 (4): 483-496.

Sun, A; Lim, EP; Ng, WK. 2003. Performance measurement framework for hierarchical text classification. JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE AND TECHNOLOGY 54 (11): 1014-1028.
Sun, AX; Lim, EP; Ng, WK; Srivastava, J. 2004. Blocking reduction strategies in hierarchical text classification. IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING 16 (10): 1305-1308.
Sun, Q; Schommer, C; Lang, A. 2004. Integration of manual and automatic text categorization. A categorization workbench for text-based email and spam. KI 2004: ADVANCES IN ARTIFICIAL INTELLIGENCE, PROCEEDINGS 3238: 156-167. LECTURE NOTES IN COMPUTER SCIENCE
Sutcliffe, AG; Ennis, M; Hu, J. 2000. Evaluating the effectiveness of visual user interfaces for information retrieval. INTERNATIONAL JOURNAL OF HUMAN-COMPUTER STUDIES 53 (5): 741-763.
SWANSON, DR. 1986. FISH OIL, RAYNAUDS SYNDROME, AND UNDISCOVERED PUBLIC KNOWLEDGE. PERSPECTIVES IN BIOLOGY AND MEDICINE 30 (1): 7-18.
SWANSON, DR. 1986. NEW HORIZONS IN PSYCHOANALYSIS - TREATMENT OF NECROSISTIC PERSONALITY-DISORDERS. PERSPECTIVES IN BIOLOGY AND MEDICINE 29 (4): 493-498.
SWANSON, DR. 1986. SUBJECTIVE VERSUS OBJECTIVE RELEVANCE IN BIBLIOGRAPHIC RETRIEVAL-SYSTEMS. LIBRARY QUARTERLY 56 (4): 389-398.
SWANSON, DR. 1986. UNDISCOVERED PUBLIC KNOWLEDGE. LIBRARY QUARTERLY 56 (2): 103-118.
SWANSON, DR. 1987. 2 MEDICAL LITERATURES THAT ARE LOGICALLY BUT NOT BIBLIOGRAPHICALLY CONNECTED. JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE 38 (4): 228-233.
SWANSON, DR. 1990. MEDICAL LITERATURE AS A POTENTIAL SOURCE OF NEW KNOWLEDGE. BULLETIN OF THE MEDICAL LIBRARY ASSOCIATION 78 (1): 29-37.
SWANSON, DR. 1993. INTERVENING IN THE LIFE-CYCLES OF SCIENTIFIC KNOWLEDGE. LIBRARY TRENDS 41 (4): 606-631.
Swanson, DR; Smalheiser, NR. 1997. An interactive system for finding complementary literatures: A stimulus to scientific discovery. ARTIFICIAL INTELLIGENCE 91 (2): 183-203.

Swanson, DR; Smalheiser, NR. 1999. Implicit text linkages between Medline records: Using Arrowsmith as an aid to scientific discovery. LIBRARY TRENDS 48 (1): 48-59.
Swanson, DR; Smalheiser, NR; Bookstein, A. 2001. Information discovery from complementary literatures: Categorizing viruses as potential weapons. JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE AND TECHNOLOGY 52 (10): 797-812.
Tai, XY; Ren, F; Kita, K. 2002. An information retrieval model based on vector space method by supervised learning. INFORMATION PROCESSING & MANAGEMENT 38 (6): 749-764.
Takahashi, S; Takahashi, H; Tsuda, K. 2004. An efficient learning system for knowledge of asset management. KNOWLEDGE-BASED INTELLIGENT INFORMATION AND ENGINEERING SYSTEMS, PT 1, PROCEEDINGS 3213: 494-500. LECTURE NOTES IN COMPUTER SCIENCE
Takahashi, S; Takahashi, M; Takahashi, H; Tsuda, K. 2005. Learning value-added information of asset management from analyst reports through text mining. KNOWLEDGE-BASED INTELLIGENT INFORMATION AND ENGINEERING SYSTEMS, PT 4, PROCEEDINGS 3684: 785-791. LECTURE NOTES IN ARTIFICIAL INTELLIGENCE
Takama, Y; Hirota, K. 2003. Web information visualization method employing immune network model for finding topic stream from document-set sequence. NEW GENERATION COMPUTING 21 (1): 49-59.
Takamura, H; Okumura, M. 2005. A comparative study on the use of labeled and unlabeled data for large margin classifiers. NATURAL LANGUAGE PROCESSING - IJCNLP 2004 3248: 456-465. LECTURE NOTES IN COMPUTER SCIENCE
Takano, A; Niwa, Y; Nishioka, S; Iwayama, M; Hisamitsu, T; Imaichi, O; Sakurai, H. 2000. Information access based on associative calculation. SOFSEM 2000: THEORY AND PRACTICE OF INFORMATICS 1963: 187-201. LECTURE NOTES IN COMPUTER SCIENCE
Takasu, A; Tanaka, K. 2004. Feature word tracking in time series documents. INTELLIGENT DATA ENGINEERING AND AUTOMATED LEARNING IDEAL 2004, PROCEEDINGS 3177: 660-665. LECTURE NOTES IN COMPUTER SCIENCE
Takci, H; Sogukpinar, L. 2004. Centroid-based language identification using letter feature set. COMPUTATIONAL LINGUISTICS AND INTELLIGENT TEXT PROCESSING 2945: 640-648. LECTURE NOTES IN COMPUTER SCIENCE

Takeuchi, K; Collier, N. 2005. Bio-medical entity extraction using support vector machines. ARTIFICIAL INTELLIGENCE IN MEDICINE 33 (2): 125-137.
Tan, CKY. 2004. Text classification using belief augmented frames. PRICAI 2004: TRENDS IN ARTIFICIAL INTELLIGENCE, PROCEEDINGS 3157: 515-523. LECTURE NOTES IN ARTIFICIAL INTELLIGENCE
Tan, CM; Wang, YF; Lee, CD. 2002. The use of bigrams to enhance text categorization. INFORMATION PROCESSING & MANAGEMENT 38 (4): 529-546.
Tan, SB. 2005. Binary k-nearest neighbor for text categorization. ONLINE INFORMATION REVIEW 29 (4): 391-399.
Tan, SB. 2005. Neighbor-weighted K-nearest neighbor for unbalanced text corpus. EXPERT SYSTEMS WITH APPLICATIONS 28 (4): 667-671.
Tanabe, L; Scherf, U; Smith, LH; Lee, JK; Hunter, L; Weinstein, JN. 1999. MedMiner: An Internet text-mining tool for biomedical information, with application to gene expression profiling. BIOTECHNIQUES 27 (6): 1210-+.
Tanaka, H; Kumano, T; Uratani, N; Ehara, T. 1999. An efficient document clustering algorithm and its application to a document browser. INFORMATION PROCESSING & MANAGEMENT 35 (4): 541-557.
Tang, B; Shepherd, M; Heywood, MI; Luo, X. 2005. Comparing dimension reduction techniques for document clustering. ADVANCES IN ARTIFICIAL INTELLIGENCE, PROCEEDINGS 3501: 292-296. LECTURE NOTES IN COMPUTER SCIENCE
Tang, CQ; Xu, ZC; Dwarkadas, S. 2003. Peer-to-peer information retrieval using self-organizing semantic overlay networks. COMPUTER COMMUNICATION REVIEW 33 (4): 175-186.
Tang, J; Li, JZ; Wang, KH; Cai, YR. 2004. Loss minimization based keyword distillation. ADVANCED WEB TECHNOLOGIES AND APPLICATIONS 3007: 572-577. LECTURE NOTES IN COMPUTER SCIENCE
Tang, N; Vemuri, VR. 2005. User-interest-based document filtering via semi-supervised clustering. FOUNDATIONS OF INTELLIGENT SYSTEMS, PROCEEDINGS 3488: 573-582. LECTURE NOTES IN COMPUTER SCIENCE
Tang, XJ; Liu, YJ; Zhang, W. 2005. Computerized support for idea generation during knowledge creating process. KNOWLEDGE-BASED INTELLIGENT INFORMATION AND ENGINEERING SYSTEMS, PT 4, PROCEEDINGS 3684: 437-443. LECTURE NOTES IN ARTIFICIAL INTELLIGENCE

Tauritz, DR; Sprinkhuizen-Kuyper, IG. 1999. Adaptive information filtering algorithms. ADVANCES IN INTELLIGENT DATA ANALYSIS, PROCEEDINGS 1642: 513-524. LECTURE NOTES IN COMPUTER SCIENCE
Terada, A; Tokunaga, T. 2003. Corpus based method of transforming nominalized phrases into clauses for text mining application. IEICE TRANSACTIONS ON INFORMATION AND SYSTEMS E86D (9): 1736-1744.
Tezuka, T; Tanaka, K. 2005. Landmark extraction: A web mining approach. SPATIAL INFORMATION THEORY, PROCEEDINGS 3693: 379-396. LECTURE NOTES IN COMPUTER SCIENCE
Theeramunkong, T. 2004. Applying passage in Web text mining. INTERNATIONAL JOURNAL OF INTELLIGENT SYSTEMS 19 (1-2): 149-158.
Thelwall, M; Wilkinson, D. 2004. Finding similar academic Web sites with links, bibliometric couplings and colinks. INFORMATION PROCESSING & MANAGEMENT 40 (3): 515-526.
Tho, QT; Hui, SC; Fong, A. 2003. Web mining for identifying research trends. DIGITAL LIBRARIES: TECHNOLOGY AND MANAGEMENT OF INDIGENOUS KNOWLEDGE FOR GLOBAL ACCESS 2911: 290-301. LECTURE NOTES IN COMPUTER SCIENCE
Thomas, JC; Kellogg, WA; Erickson, T. 2001. The knowledge management puzzle: Human and social factors in knowledge management. IBM SYSTEMS JOURNAL 40 (4): 863-884.
Thompson, P. 2005. Text mining, names and security. JOURNAL OF DATABASE MANAGEMENT 16 (1): 54-59.
Tiffin, N; Kelso, JF; Powell, AR; Pan, H; Bajic, VB; Hide, WA. 2005. Integration of text- and data-mining using ontologies successfully selects disease gene candidates. NUCLEIC ACIDS RESEARCH 33 (5): 1544-1552.
Tikk, D; Yang, JD; Bang, SL. 2003. Hierarchical text categorization using fuzzy relational thesaurus. KYBERNETIKA 39 (5): 583-600.
Toivonen, J; Visa, A; Vesanen, T; Back, B; Vanharanta, H. 2001. Validation of text clustering based on document contents. MACHINE LEARNING AND DATA MINING IN PATTERN RECOGNITION 2123: 184-195. LECTURE NOTES IN ARTIFICIAL INTELLIGENCE
Tombros, A; Jose, JM; Ruthven, I. 2003. Clustering top-ranking sentences for information access. RESEARCH AND ADVANCED TECHNOLOGY FOR DIGITAL LIBRARIES 2769: 523-528. LECTURE NOTES IN COMPUTER SCIENCE

Tombros, A; van Rijsbergen, CJ. 2004. Query-sensitive similarity measures for information retrieval. KNOWLEDGE AND INFORMATION SYSTEMS 6 (5): 617-642.
Tombros, A; Villa, R; Van Rijsbergen, CJ. 2002. The effectiveness of query-specific hierarchic clustering in information retrieval. INFORMATION PROCESSING & MANAGEMENT 38 (4): 559-582.
Tomsich, P; Rauber, A; Merkl, D. 2000. parSOM: Using parallelism to overcome memory latency in self-organizing neural networks. HIGH PERFORMANCE COMPUTING AND NETWORKING, PROCEEDINGS 1823: 136-145. LECTURE NOTES IN COMPUTER SCIENCE
Tong, S; Koller, D. 2002. Support vector machine active learning with applications to text classification. JOURNAL OF MACHINE LEARNING RESEARCH 2 (1): 45-66.
Topchy, A; Punch, W. 2003. Dimensionality reduction via genetic value clustering. GENETIC AND EVOLUTIONARY COMPUTATION - GECCO 2003, PT II, PROCEEDINGS 2724: 1431-1443. LECTURE NOTES IN COMPUTER SCIENCE
Torkkola, K. 2003. Discriminative features for text document classification. PATTERN ANALYSIS AND APPLICATIONS 6 (4): 301-308.
Torvik, VI; Weeber, M; Swanson, DR; Smalheiser, NR. 2005. A probabilistic similarity metric for Medline records: A model for author name disambiguation. JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE AND TECHNOLOGY 56 (2): 140-158.
Toselli, AH; Pastor, M; Juan, A; Vidal, E. 2005. Spontaneous handwriting text recognition and classification using finite-state models. PATTERN RECOGNITION AND IMAGE ANALYSIS, PT 2, PROCEEDINGS 3523: 363-370. LECTURE NOTES IN COMPUTER SCIENCE
Trybula, WJ. 1999. Text mining. ANNUAL REVIEW OF INFORMATION SCIENCE AND TECHNOLOGY 34: 385-419.
Tsay, JJ; Wang, JD. 2004. Improving linear classifier for Chinese text categorization. INFORMATION PROCESSING & MANAGEMENT 40 (2): 223-237.
Tseng, CM; Tsai, KH; Hsu, CC; Chang, HC. 2005. On the Chinese document clustering based on dynamical term clustering. INFORMATION RETRIEVAL TECHNOLOGY, PROCEEDINGS 3689: 534-539. LECTURE NOTES IN COMPUTER SCIENCE
Tseng, FSC; Hwung, WJ. 2002. An automatic load/extract scheme for XML documents through object-relational repositories. JOURNAL OF SYSTEMS AND SOFTWARE 64 (3): 207-218.

Turney, PD; Littman, ML. 2003. Measuring praise and criticism: Inference of semantic orientation from association. ACM TRANSACTIONS ON INFORMATION SYSTEMS 21 (4): 315-346.
Uramoto, N; Matsuzawa, H; Nagano, T; Murakami, A; Takeuchi, H; Takeda, K. 2004. A text-mining system for knowledge discovery from Biomedical Documents. IBM SYSTEMS JOURNAL 43 (3): 516-533.
Urena-Lopez, LA; Buenaga, M; Gomez, JM. 2001. Integrating linguistic resources in TC through WSD. COMPUTERS AND THE HUMANITIES 35 (2): 215-230.
van der Eijk, CC; van Mulligen, EM; Kors, JA; Mons, B; van den Berg, J. 2004. Constructing an associative concept space for literature-based discovery. JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE AND TECHNOLOGY 55 (5): 436-444.
VANRIJSBERGEN, CJ; CROFT, WB. 1975. DOCUMENT CLUSTERING - EVALUATION OF SOME EXPERIMENTS WITH CRANFIELD 1400 COLLECTION. INFORMATION PROCESSING & MANAGEMENT 11 (5-7): 171-182.
Vembu, S; Baumann, S. 2005. A self-organizing map based knowledge discovery for music recommendation systems. COMPUTER MUSIC MODELING AND RETRIEVAL 3310: 119-129. LECTURE NOTES IN COMPUTER SCIENCE
Vert, JP. 2001. Adaptive context trees and text clustering. IEEE TRANSACTIONS ON INFORMATION THEORY 47 (5): 1884-1901.
Vert, JP. 2001. Text categorization using adaptive context trees. COMPUTATIONAL LINGUISTICS AND INTELLIGENT TEXT PROCESSING 2004: 423-436. LECTURE NOTES IN COMPUTER SCIENCE
Viator, JA; Pestorius, FM. 2001. Investigating trends in acoustics research from 1970-1999. JOURNAL OF THE ACOUSTICAL SOCIETY OF AMERICA 109 (5): 1779-1783, Part 1.
Vilar, D; Castro, MJ; Sanchis, E. 2004. Multi-label text classification using multinomial models. ADVANCES IN NATURAL LANGUAGE PROCESSING 3230: 220-230. LECTURE NOTES IN ARTIFICIAL INTELLIGENCE
Vinciarelli, A. 2005. Noisy text categorization. IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE 27 (12): 1882-1895.
Vinokourov, A; Girolami, M. 2002. A probabilistic framework for the hierarchic organisation and classification of document collections.

JOURNAL OF INTELLIGENT INFORMATION SYSTEMS 18 (2-3): 153-172.
Vinot, R; Yvon, F. 2003. Improving Rocchio with weakly supervised clustering. MACHINE LEARNING: ECML 2003 2837: 456-467. LECTURE NOTES IN ARTIFICIAL INTELLIGENCE
Visa, A. 2001. Technology of text mining. MACHINE LEARNING AND DATA MINING IN PATTERN RECOGNITION 2123: 1-11. LECTURE NOTES IN ARTIFICIAL INTELLIGENCE
Visa, A; Toivonen, J; Vanharanta, H; Back, B. 2002. Contents matching defined by prototypes: Methodology verification with books of the bible. JOURNAL OF MANAGEMENT INFORMATION SYSTEMS 18 (4): 87-100.
Vittaut, JN; Amini, MR; Gallinari, P. 2002. Learning classification with both labeled and unlabeled data. MACHINE LEARNING: ECML 2002 2430: 468-479. LECTURE NOTES IN ARTIFICIAL INTELLIGENCE
Voss, A; Althoff, KD; Becker-Kornstaedt, U; Decker, B; Klotz, A; Leopold, E; Rech, J. 2002. Enhancing experience management and process learning with moderated discourses: The indiGo approach. PRACTICAL ASPECTS OF KNOWLEDGE MANAGEMENT 2569: 114-125. LECTURE NOTES IN ARTIFICIAL INTELLIGENCE
Walter, JA. 2004. H-MDS: a new approach for interactive visualization with multidimensional scaling in the hyperbolic space. INFORMATION SYSTEMS 29 (4): 273-292.
Wang, C; Wang, WY. 2005. Using term clustering and supervised term affinity construction to boost text classification. ADVANCES IN KNOWLEDGE DISCOVERY AND DATA MINING, PROCEEDINGS 3518: 813-819. LECTURE NOTES IN ARTIFICIAL INTELLIGENCE
Wang, DL; Yu, G; Bao, YB; Zhang, M. 2005. An optimized K-Means algorithm of reducing cluster intra-dissimilarity for document clustering. ADVANCES IN WEB-AGE INFORMATION MANAGEMENT, PROCEEDINGS 3739: 785-790. LECTURE NOTES IN COMPUTER SCIENCE
Wang, HY; Chen, Y; Dai, YQ. 2005. A soft real-time web news classification system with double control loops. ADVANCES IN WEB-AGE INFORMATION MANAGEMENT, PROCEEDINGS 3739: 81-90. LECTURE NOTES IN COMPUTER SCIENCE
Wang, JTL; Liu, JH; Wang, JH. 2005. XML clustering and retrieval through principal component analysis. INTERNATIONAL JOURNAL ON ARTIFICIAL INTELLIGENCE TOOLS 14 (4): 683-699.

Wang, K; Liu, HQ. 2000. Discovering structural association of semistructured data. IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING 12 (3): 353-371.
Wang, Q; Ng, YK. 2003. An ontology-based binary-categorization approach for recognizing multiple-record web documents using a probabilistic retrieval model. INFORMATION RETRIEVAL 6 (3-4): 295-332.
Wang, SY; Yu, L; Lai, KK. 2004. A novel hybrid AI system framework for crude oil price forecasting. DATA MINING AND KNOWLEDGE MANAGEMENT 3327: 233-242. LECTURE NOTES IN ARTIFICIAL INTELLIGENCE
Wang, W; Do, DB; Lin, XM. 2005. Term graph model for text classification. ADVANCED DATA MINING AND APPLICATIONS, PROCEEDINGS 3584: 19-30. LECTURE NOTES IN ARTIFICIAL INTELLIGENCE
Wang, YT; Kitsuregawa, M. 2001. Link based clustering of Web search results. ADVANCES IN WEB-AGE INFORMATION MANAGEMENT, PROCEEDINGS 2118: 225-236. LECTURE NOTES IN COMPUTER SCIENCE
Watts, RJ; Porter, AL; Cunningham, S; Zhu, DH. 1997. TOAS intelligence mining; Analysis of natural language processing and computational linguistics. PRINCIPLES OF DATA MINING AND KNOWLEDGE DISCOVERY 1263: 323-334. LECTURE NOTES IN ARTIFICIAL INTELLIGENCE
Wee, LKA; Tong, LC; Tan, CL. 1998. Knowledge representation issues in information extraction. PRICAI'98: TOPICS IN ARTIFICIAL INTELLIGENCE 1531: 448-458. LECTURE NOTES IN ARTIFICIAL INTELLIGENCE
Weeber, M; Klein, H; Aronson, AR; Mork, JG; de Jong-van den Berg, LTW; Vos, R. 2000. Text-based discovery in biomedicine: The architecture of the DAD-system. JOURNAL OF THE AMERICAN MEDICAL INFORMATICS ASSOCIATION: 903-907, Suppl. S.
Weeber, M; Klein, H; de Jong-van den Berg, LTW; Vos, R. 2001. Using concepts in literature-based discovery: Simulating Swanson's Raynaud-fish oil and migraine-magnesium discoveries. JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE AND TECHNOLOGY 52 (7): 548-557.
Weeber, M; Kors, JA; Mons, B. 2005. Online tools to support literature-based discovery in the life sciences. BRIEFINGS IN BIOINFORMATICS 6 (3): 277-286.

Wei, CP; Cheng, TH; Pai, YC. 2006. Semantic enrichment in knowledge repositories: Annotating semantic relationships between discussion documents. JOURNAL OF DATABASE MANAGEMENT 17 (1): 49-66.
Wei, CP; Hu, PJ; Dong, YX. 2002. Managing document categories in e-commerce environments: an evolution-based approach. EUROPEAN JOURNAL OF INFORMATION SYSTEMS 11 (3): 208-222.
Wei, CP; Lee, YH. 2004. Event detection from Online news documents for supporting environmental scanning. DECISION SUPPORT SYSTEMS 36 (4): 385-401.
Wei, CP; Yang, CS; Hsiao, HW; Cheng, TH. 2006. Combining preference- and content-based approaches for improving document clustering effectiveness. INFORMATION PROCESSING & MANAGEMENT 42 (2): 350-372.
Weiss, SM; Apte, C; Damerau, FJ; Johnson, DE; Oles, FJ; Goetz, T; Hampp, T. 1999. Maximizing text-mining performance. IEEE INTELLIGENT SYSTEMS & THEIR APPLICATIONS 14 (4): 63-69.
Weiss, SM; Apte, CV. 2002. Automated generation of model cases for help-desk applications. IBM SYSTEMS JOURNAL 41 (3): 421-427.
Weng, SS; Lin, YJ. 2003. A study on searching for similar documents based on multiple concepts and distribution of concepts. EXPERT SYSTEMS WITH APPLICATIONS 25 (3): 355-368.
Weng, SS; Liu, CK. 2004. Using text classification and multiple concepts to answer e-mails. EXPERT SYSTEMS WITH APPLICATIONS 26 (4): 529-543.
Wermter, S. 2000. Neural network agents for learning semantic text classification. INFORMATION RETRIEVAL 3 (2): 87-103.
Wiebe, J; Wilson, T; Bruce, R; Bell, M; Martin, M. 2004. Learning subjective language. COMPUTATIONAL LINGUISTICS 30 (3): 277-308.
WILBUR, WJ. 1992. RETRIEVAL TESTING BY THE COMPARISON OF STATISTICALLY INDEPENDENT RETRIEVAL METHODS. JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE 43 (5): 358-370.
Wilbur, WJ; Yang, YM. 1996. An analysis of statistical term strength and its use in the indexing and retrieval of molecular biology texts. COMPUTERS IN BIOLOGY AND MEDICINE 26 (3): 209-222.
WILLETT, P. 1980. DOCUMENT CLUSTERING USING AN INVERTED FILE APPROACH. JOURNAL OF INFORMATION SCIENCE 2 (5): 223-231.
Williamson, J; Dooley, K; Corman, S. 2004. Using text mining to create actionable knowledge: Application to network failure incident reports.

PRACTICAL ASPECTS OF KNOWLEDGE MANAGEMENT, PROCEEDINGS 3336: 71-81. LECTURE NOTES IN ARTIFICIAL INTELLIGENCE
Wiratunga, N; Lothian, R; Chakraborti, S; Koychev, I. 2005. A propositional approach to textual case indexing. KNOWLEDGE DISCOVERY IN DATABASES: PKDD 2005 3721: 380-391. LECTURE NOTES IN ARTIFICIAL INTELLIGENCE
Wise, MJ. 2000. Protein Annotators' Assistant: A novel application of information retrieval techniques. JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE 51 (12): 1131-1136.
Witte, R; Baker, CJO. 2005. Combining biological databases and text mining to support new bioinformatics applications. NATURAL LANGUAGE PROCESSING AND INFORMATION SYSTEMS, PROCEEDINGS 3513: 310-321. LECTURE NOTES IN COMPUTER SCIENCE
Witten, IH. 2000. Browsing around a digital library: Today and tomorrow. COMBINATORIAL PATTERN MATCHING 1848: 12-26. LECTURE NOTES IN COMPUTER SCIENCE
Witten, IH. 2002. Learning structure from sequences, with applications in a digital library. ALGORITHMIC LEARNING THEORY, PROCEEDINGS 2533: 42-56. LECTURE NOTES IN ARTIFICIAL INTELLIGENCE
Witter, DI; Berry, MW. 1998. Downdating the latent semantic indexing model for conceptual information retrieval. COMPUTER JOURNAL 41 (8): 589-601.
Wnek, J. 2005. LSI-based taxonomy generation: The taxonomist system. INTELLIGENCE AND SECURITY INFORMATICS, PROCEEDINGS 3495: 389-394. LECTURE NOTES IN COMPUTER SCIENCE
Wong, TL; Lam, W; Wang, W. 2003. Beyond supervised learning of wrappers for extracting information from unseen Web sites. INTELLIGENT DATA ENGINEERING AND AUTOMATED LEARNING 2690: 725-733. LECTURE NOTES IN COMPUTER SCIENCE
Wormell, I. 2000. Bibliometric analysis of the Welfare Topic. SCIENTOMETRICS 48 (2): 203-236.
Wormell, I. 2000. Critical aspects of the Danish Welfare State - as revealed by issue tracking. SCIENTOMETRICS 48 (2): 237-250.
Wren, JD. 2006. Using fuzzy set theory and scale-free network properties to relate MEDLINE terms. SOFT COMPUTING 10 (4): 374-381.
Wu, JQ; Wu, YZ; Liu, J; Zhuang, YT. 2004. Multi-document summarization based on link analysis and text classification. DIGITAL LIBRARIES: INTERNATIONAL COLLABORATION AND CROSS-FERTILIZATION,

PROCEEDINGS 3334: 649-649. LECTURE NOTES IN COMPUTER SCIENCE
Wu, XY; Srihari, R; Zheng, ZH. 2004. Document representation for one-class SVM. MACHINE LEARNING: ECML 2004, PROCEEDINGS 3201: 489-500. LECTURE NOTES IN COMPUTER SCIENCE
Wu, ZH; Zhou, XZ; Liu, BY; Chen, JL. 2004. Text mining for finding functional community of related genes using TCM knowledge. KNOWLEDGE DISCOVERY IN DATABASES: PKDD 2004, PROCEEDINGS 3202: 459-470. LECTURE NOTES IN ARTIFICIAL INTELLIGENCE
Xu, ST; Zhang, J. 2004. A parallel hybrid web document clustering algorithm and its performance study. JOURNAL OF SUPERCOMPUTING 30 (2): 117-131.
Xu, X; Zhang, BF; Zhong, QX. 2005. Text categorization using SVMs with Rocchio ensemble for Internet information classification. NETWORKING AND MOBILE COMPUTING, PROCEEDINGS 3619: 1022-1031. LECTURE NOTES IN COMPUTER SCIENCE
Xu, YH; Umemura, K. 2003. Optimal local dimension analysis of latent semantic indexing on query neighbor space. IEICE TRANSACTIONS ON INFORMATION AND SYSTEMS E86D (9): 1762-1772.
Xu, Z; Yu, K; Tresp, V; Xu, XW; Wang, JZ. 2003. Representative sampling for text classification using support vector machines. ADVANCES IN INFORMATION RETRIEVAL 2633: 393-407. LECTURE NOTES IN COMPUTER SCIENCE
Xue, D; Sun, MS. 2003. Chinese text categorization based on the binary weighting model with non-binary smoothing. ADVANCES IN INFORMATION RETRIEVAL 2633: 408-419. LECTURE NOTES IN COMPUTER SCIENCE
Xue, DJ; Sun, MS. 2003. A study on feature weighting in Chinese text categorization. COMPUTATIONAL LINGUISTICS AND INTELLIGENT TEXT PROCESSING, PROCEEDINGS 2588: 592-601. LECTURE NOTES IN COMPUTER SCIENCE
Xue, DJ; Sun, MS. 2004. Eliminating high-degree biased character bigrams for dimensionality reduction in Chinese text categorization. ADVANCES IN INFORMATION RETRIEVAL, PROCEEDINGS 2997: 197-208. LECTURE NOTES IN COMPUTER SCIENCE
Xue, DJ; Sun, MS. 2004. Raising high-degree overlapped character bigrams into trigrams for dimensionality reduction in Chinese text categorization. COMPUTATIONAL LINGUISTICS AND INTELLIGENT TEXT

PROCESSING 2945: 584-595. LECTURE NOTES IN COMPUTER SCIENCE
Yamamoto, S; Asanuma, T; Takagi, T; Fukuda, KI. 2004. The Molecule Role Ontology: an ontology for annotation of signal transduction pathway molecules in the scientific literature. COMPARATIVE AND FUNCTIONAL GENOMICS 5 (6-7): 528-536.
Yamasaki, M; Takeda, M; Fukuda, T; Nanri, I. 1998. Discovering characteristic patterns from collections of classical Japanese poems. DISCOVERY SCIENCE 1532: 129-140. LECTURE NOTES IN ARTIFICIAL INTELLIGENCE
Yan, W. 2005. Network attack scenarios extraction and categorization by mining IDS alert streams. JOURNAL OF UNIVERSAL COMPUTER SCIENCE 11 (8): 1367-1382.
Yan, X; Li, X; Song, DW. 2004. A correlation analysis on LSA and HAL semantic space models. COMPUTATIONAL AND INFORMATION SCIENCE, PROCEEDINGS 3314: 711-717. LECTURE NOTES IN COMPUTER SCIENCE
Yang, CF; Ye, M; Zhao, J. 2005. Document clustering based on nonnegative sparse matrix factorization. ADVANCES IN NATURAL COMPUTATION, PT 2, PROCEEDINGS 3611: 557-563. LECTURE NOTES IN COMPUTER SCIENCE
Yang, HC; Lee, CH. 2003. Building topic maps using a text mining approach. FOUNDATIONS OF INTELLIGENT SYSTEMS 2871: 307-314. LECTURE NOTES IN ARTIFICIAL INTELLIGENCE
Yang, HC; Lee, CH. 2003. Discovering image semantics from web pages using a text mining approach. ADVANCES IN WEB-AGE INFORMATION MANAGEMENT, PROCEEDINGS 2762: 495-502. LECTURE NOTES IN COMPUTER SCIENCE
Yang, HC; Lee, CH. 2003. Mining environmental texts of images in web pages for image retrieval. FOUNDATIONS OF INTELLIGENT SYSTEMS 2871: 334-338. LECTURE NOTES IN ARTIFICIAL INTELLIGENCE
Yang, HC; Lee, CH. 2004. A text mining approach on automatic generation of web directories and hierarchies. EXPERT SYSTEMS WITH APPLICATIONS 27 (4): 645-663.
Yang, HC; Lee, CH. 2005. A text mining approach for automatic construction of hypertexts. EXPERT SYSTEMS WITH APPLICATIONS 29 (4): 723-734.
Yang, HC; Lee, CH. 2005. Automatic category theme identification and hierarchy generation for Chinese text categorization. JOURNAL OF INTELLIGENT INFORMATION SYSTEMS 25 (1): 47-67.

Yang, JW; Chen, XO. 2002. A semi-structured document model for text mining. JOURNAL OF COMPUTER SCIENCE AND TECHNOLOGY 17 (5): 603-610.
Yang, L; Rahi, A. 2003. Dynamic clustering of web search results. COMPUTATIONAL SCIENCE AND ITS APPLICATIONS - ICCSA 2003, PT 1, PROCEEDINGS 2667: 153-159. LECTURE NOTES IN COMPUTER SCIENCE
Yang, YM; Carbonell, JG; Brown, RD; Frederking, RE. 1998. Translingual information retrieval: learning from bilingual corpora. ARTIFICIAL INTELLIGENCE 103 (1-2): 323-345.
YANG, YM; CHUTE, CG. 1994. AN APPLICATION OF EXPERT NETWORK TO CLINICAL CLASSIFICATION AND MEDLINE INDEXING. JOURNAL OF THE AMERICAN MEDICAL INFORMATICS ASSOCIATION: 157-161, Suppl. S.
YANG, YM; CHUTE, CG. 1994. AN EXAMPLE-BASED MAPPING METHOD FOR TEXT CATEGORIZATION AND RETRIEVAL. ACM TRANSACTIONS ON INFORMATION SYSTEMS 12 (3): 252-277.
Yang, YM; Slattery, S; Ghani, R. 2002. A study of approaches to hypertext categorization. JOURNAL OF INTELLIGENT INFORMATION SYSTEMS 18 (2-3): 219-241.
Yang, YM; Wilbur, J. 1996. Using corpus statistics to remove redundant words in text categorization. JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE 47 (5): 357-369.
Ye, JP; Janardan, R; Park, CH; Park, H. 2004. An optimization criterion for generalized discriminant analysis on undersampled problems. IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE 26 (8): 982-994.
Ye, JP; Li, Q. 2005. A two-stage linear discriminant analysis via QR-decomposition. IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE 27 (6): 929-941.
Yin, ZH; Wang, YC; Song, JP; Zeng, YM. 2002. Filtering multi-category noise in knowledge base for automatic text classification. CHINESE JOURNAL OF ELECTRONICS 11 (4): 450-453.
Yoon, B; Park, Y. 2005. A systematic approach for identifying technology opportunities: Keyword-based morphology analysis. TECHNOLOGICAL FORECASTING AND SOCIAL CHANGE 72 (2): 145-160.
Yoon, JP; Raghavan, V; Chakilam, V; Kerschberg, L. 2001. BitCube: A three-dimensional bitmap indexing for XML documents. JOURNAL OF INTELLIGENT INFORMATION SYSTEMS 17 (2-3): 241-254.

Yoon, Y; Lee, C; Lee, GG. 2005. Systematic construction of hierarchical classifier in SVM-based text categorization. NATURAL LANGUAGE PROCESSING - IJCNLP 2004 3248: 616-625. LECTURE NOTES IN COMPUTER SCIENCE
Yoon, Y; Lee, GG. 2005. Practical application of associative classifier for document classification. INFORMATION RETRIEVAL TECHNOLOGY, PROCEEDINGS 3689: 467-478. LECTURE NOTES IN COMPUTER SCIENCE
Yu, S; Song, H; Ma, FY. 2004. Novel SVM performance estimators for information retrieval systems. ADVANCED WEB TECHNOLOGIES AND APPLICATIONS 3007: 895-898. LECTURE NOTES IN COMPUTER SCIENCE
Yuan, ST; Sun, J. 2005. Ontology-based structured cosine similarity in document summarization: With applications to mobile audio-based knowledge management. IEEE TRANSACTIONS ON SYSTEMS MAN AND CYBERNETICS PART B-CYBERNETICS 35 (5): 1028-1040.
Yukari, I; Satoru, T; Kazuhiko, T. 2004. A study of knowledge extraction from free text data in customer satisfaction survey. KNOWLEDGE-BASED INTELLIGENT INFORMATION AND ENGINEERING SYSTEMS, PT 1, PROCEEDINGS 3213: 509-515. LECTURE NOTES IN COMPUTER SCIENCE
Zamir, O; Etzioni, O. 1999. Grouper: a dynamic clustering interface to Web search results. COMPUTER NETWORKS-THE INTERNATIONAL JOURNAL OF COMPUTER AND TELECOMMUNICATIONS NETWORKING 31 (11-16): 1361-1374.
Zelikovitz, S; Hirsh, H. 2002. Integrating background knowledge into nearest-neighbor text classification. ADVANCES IN CASE-BASED REASONING 2416: 1-5. LECTURE NOTES IN ARTIFICIAL INTELLIGENCE
Zelikovitz, S; Marquez, F. 2005. Transductive learning for short-text classification problems using latent semantic indexing. INTERNATIONAL JOURNAL OF PATTERN RECOGNITION AND ARTIFICIAL INTELLIGENCE 19 (2): 143-163.
Zha, HY; Simon, HD. 1999. On updating problems in latent semantic indexing. SIAM JOURNAL ON SCIENTIFIC COMPUTING 21 (2): 782-791.
Zha, HY; Zhang, ZY. 2000. Matrices with low-rank-plus-shift structure: Partial SVD and latent semantic indexing. SIAM JOURNAL ON MATRIX ANALYSIS AND APPLICATIONS 21 (2): 522-536.

Zhang, BC; Chen, X; Gao, W. 2005. Discriminant analysis based on kernelized decision boundary for face recognition. AUDIO AND VIDEO BASED BIOMETRIC PERSON AUTHENTICATION, PROCEEDINGS 3546: 966-976. LECTURE NOTES IN COMPUTER SCIENCE
Zhang, J. 2001. The characteristic analysis of the DARE visual space. INFORMATION RETRIEVAL 4 (1): 61-78.
Zhang, J; Chen, XY; Chen, Y; Hui, YF. 2005. Association classification based on sample weighting. FUZZY SYSTEMS AND KNOWLEDGE DISCOVERY, PT 2, PROCEEDINGS 3614: 624-633, Part 2. LECTURE NOTES IN ARTIFICIAL INTELLIGENCE
Zhang, M; Yang, DQ; Deng, ZH; Feng, Y; Wang, WQ; Zhao, PX; Wu, S; Wang, SA; Tang, SW. 2004. PKUSpace: A collaborative platform for scientific researching. ADVANCES IN WEB-BASED LEARNING - ICWL 2004 3143: 120-127. LECTURE NOTES IN COMPUTER SCIENCE
Zhang, MY; Lu, ZD; Zou, CY. 2004. A Chinese word segmentation based on language situation in processing ambiguous words. INFORMATION SCIENCES 162 (3-4): 275-285.
Zhang, T; Iyengar, VS. 2002. Recommender systems using linear classifiers. JOURNAL OF MACHINE LEARNING RESEARCH 2 (3): 313-334.
Zhang, T; Oles, FJ. 2001. Text categorization based on regularized linear classification methods. INFORMATION RETRIEVAL 4 (1): 5-31.
Zhang, X; Zhu, XY. 2005. Extended Bi-gram features in text categorization. PATTERN RECOGNITION AND IMAGE ANALYSIS, PT 2, PROCEEDINGS 3523: 379-386. LECTURE NOTES IN COMPUTER SCIENCE
Zhang, XY; Berry, MW; Raghavan, P. 2001. Level search schemes for information filtering and retrieval. INFORMATION PROCESSING & MANAGEMENT 37 (2): 313-334.
Zhang, Y; Zhang, LJ; Li, ZH; Yan, JF. 2003. Improving the performance of text classifiers by using association features. FOUNDATIONS OF INTELLIGENT SYSTEMS 2871: 315-319. LECTURE NOTES IN ARTIFICIAL INTELLIGENCE
Zhang, Z; Zhou, SG; Zhou, AY. 2004. Sequential classifiers combination for text categorization: An experimental study. ADVANCES IN WEB-AGE INFORMATION MANAGEMENT: PROCEEDINGS 3129: 509-518. LECTURE NOTES IN COMPUTER SCIENCE
Zhang, ZH; Shen, H. 2005. Application of online-training SVMs for real-time intrusion detection with different considerations. COMPUTER COMMUNICATIONS 28 (12): 1428-1442.

Zhao, H; Lu, BL. 2004. Modular k-nearest neighbor classification method for massively parallel text categorization. COMPUTATIONAL AND INFORMATION SCIENCE, PROCEEDINGS 3314: 867-872. LECTURE NOTES IN COMPUTER SCIENCE
Zhao, R; Grosky, WI. 2002. Narrowing the semantic gap - Improved text-based web document retrieval using visual features. IEEE TRANSACTIONS ON MULTIMEDIA 4 (2): 189-200.
Zhao, R; Grosky, WI. 2002. Negotiating the semantic gap: from feature maps to semantic landscapes. PATTERN RECOGNITION 35 (3): 593-600.
Zhao, Y; Karypis, G. 2004. Empirical and theoretical comparisons of selected criterion functions for document clustering. MACHINE LEARNING 55 (3): 311-331.
Zhao, Y; Karypis, G. 2005. Hierarchical clustering algorithms for document datasets. DATA MINING AND KNOWLEDGE DISCOVERY 10 (2): 141-168.
Zhdanova, AV; Shishkin, DV. 2002. Classification of email queries by topic: Approach based on hierarchically structured subject domain. INTELLIGENT DATA ENGINEERING AND AUTOMATED LEARNING - IDEAL 2002 2412: 99-104. LECTURE NOTES IN COMPUTER SCIENCE
Zheng, XS; Liu, WL; He, PL; Da, WD. 2004. Document clustering algorithm based on tree-structured growing self-organizing feature map. ADVANCES IN NEURAL NETWORKS - ISNN 2004, PT 1 3173: 840-845. LECTURE NOTES IN COMPUTER SCIENCE
Zhong, N; Matsunaga, T; Liu, CN. 2002. A text mining agents based architecture for personal e-mail filtering and management. INTELLIGENT DATA ENGINEERING AND AUTOMATED LEARNING - IDEAL 2002 2412: 329-336. LECTURE NOTES IN COMPUTER SCIENCE
Zhong, S. 2005. Efficient streaming text clustering. NEURAL NETWORKS 18 (5-6): 790-798.
Zhong, S; Ghosh, J. 2005. Generative model-based document clustering: a comparative study. KNOWLEDGE AND INFORMATION SYSTEMS 8 (3): 374-384.
Zhou, AY; Qian, WI; Qian, HL; Zhang, L; Liang, YQ; Jin, W. 2002. Clustering DTDs: An interactive two-level approach. JOURNAL OF COMPUTER SCIENCE AND TECHNOLOGY 17 (6): 807-819.
Zhou, SG; Guan, JH. 2002. Evaluation and construction of training corpuses for text classification: A preliminary study. NATURAL LANGUAGE PROCESSING AND INFORMATION SYSTEMS 2553: 97-108. LECTURE NOTES IN COMPUTER SCIENCE

Zhou, XD; Wang, T; Zhou, HP; Chen, HW. 2004. Categorizing Web information on subject with statistical language modeling. WEB INFORMATION SYSTEMS - WISE 2004, PROCEEDINGS 3306: 403-408. LECTURE NOTES IN COMPUTER SCIENCE
Zhou, XZ; Wu, ZH. 2004. Distributional character clustering for chinese text categorization. PRICAI 2004: TRENDS IN ARTIFICIAL INTELLIGENCE, PROCEEDINGS 3157: 575-584. LECTURE NOTES IN ARTIFICIAL INTELLIGENCE
Zhou, ZH; Jiang, K; Li, M. 2005. Multi-instance learning based web mining. APPLIED INTELLIGENCE 22 (2): 135-147.
Zhu, DH; Porter, AL. 2002. Automated extraction and visualization of information for technological intelligence and forecasting. TECHNOLOGICAL FORECASTING AND SOCIAL CHANGE 69 (5): 495-506.
Zhu, JB; Chen, WL. 2005. Improving text categorization using domain knowledge. NATURAL LANGUAGE PROCESSING AND INFORMATION SYSTEMS, PROCEEDINGS 3513: 103-113. LECTURE NOTES IN COMPUTER SCIENCE
Zhu, JB; Chen, WL; Yao, TS. 2004. Using seed words to learn to categorize Chinese text. ADVANCES IN NATURAL LANGUAGE PROCESSING 3230: 464-473. LECTURE NOTES IN ARTIFICIAL INTELLIGENCE
Zhu, X; Cao, HL; Yu, Y. 2006. SDQE: towards automatic semantic query optimization in P2P systems. INFORMATION PROCESSING & MANAGEMENT 42 (1): 222-236.
Zhuang, L; Dai, HH; Hang, XS. 2005. A novel field learning algorithm for dual imbalance text classification. FUZZY SYSTEMS AND KNOWLEDGE DISCOVERY, PT 2, PROCEEDINGS 3614: 39-48, Part 2. LECTURE NOTES IN ARTIFICIAL INTELLIGENCE
Zu, GW; Murata, M; Ohyama, W; Wakabayashi, T; Kimura, F. 2004. The impact of OCR accuracy on automatic text classification. CONTENT COMPUTING, PROCEEDINGS 3309: 403-409. LECTURE NOTES IN COMPUTER SCIENCE
Zu, GW; Ohyama, W; Wakabayashi, T; Kimura, F. 2005. Automatic text classification of English newswire articles based on statistical classification techniques. ELECTRICAL ENGINEERING IN JAPAN 152 (1): 50-60.